**基于Hadoop离线日志分析系统**

[**项目背景介绍**](https://blog.csdn.net/weixin_41605937/article/details/107507048#%E9%A1%B9%E7%9B%AE%E8%83%8C%E6%99%AF%E4%BB%8B%E7%BB%8D)

在大数据的今天，数据已经成为我们的一部分中的重要的资源。在通过对用户的历史数据分析，可以实现对公司和企业的产品的实时分析和评价、在风控管理和实时监控、精确广告投放、可以更好的提供用户个性化的服务……。例如在一下领域的作用：

1黑匣子数据:这是飞机生成的数据，包括喷气式飞机和直升机。黑匣子数据包括机组人员声音、麦克风录音和飞机性能信息。

2社交媒体数据:由Twitter、Facebook、Instagram、Pinterest和谷歌+等社交媒体网站开发的数据。股票交易所数据:这是来自股票交易所的关于客户买卖股票决策的数据。

3电网数据:这是来自电网的数据。它包含特定节点上的信息，比如使用信息。

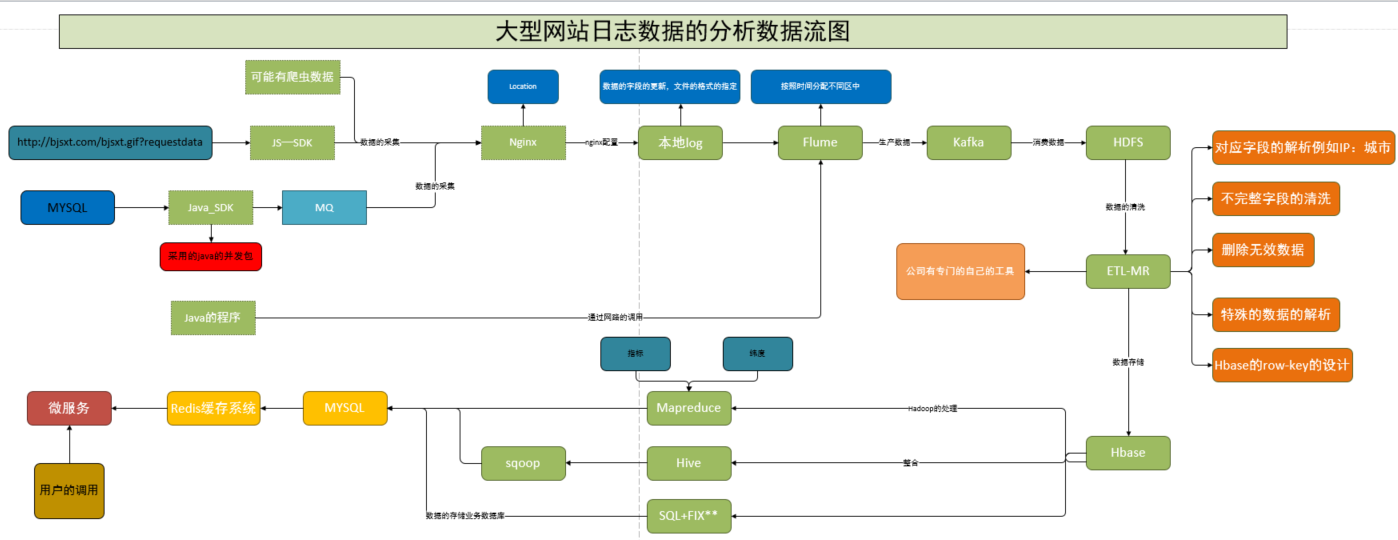
4传输数据:这包括可能的容量、车辆模型、可用性和车辆所覆盖的距离。

5搜索引擎数据:这是最大的大数据来源之一。搜索引擎拥有庞大的数据库来获取数据。

本系统的就是通对用户的浏览行为数据分析：1用户每次访问网站时所有的行为数据（访问、浏览、搜索、点击…）2 用户行为轨迹、流量日志 3 用户的操作的相关的业务的数据来构建一个大型的日志分析系统。4其他数据爬虫数据……

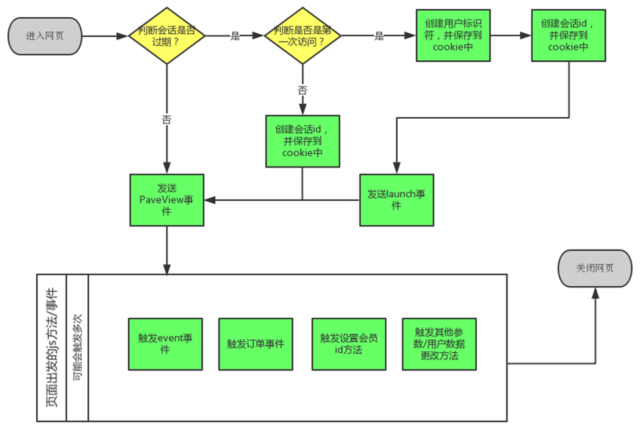
利用的hadoop实现离线的相关业务数据分析处理。 系统后期采用基于Spark Stream的实时流计算来实现的一个准实时的分析系统。

[**项目框架设计**](https://blog.csdn.net/weixin_41605937/article/details/107507048#%E9%A1%B9%E7%9B%AE%E6%A1%86%E6%9E%B6%E8%AE%BE%E8%AE%A1)



[**项目数据采集**](https://blog.csdn.net/weixin_41605937/article/details/107507048#%E9%A1%B9%E7%9B%AE%E6%95%B0%E6%8D%AE%E9%87%87%E9%9B%86)

1. **JS端的数据的采集**



针对我们最终的不同分析模块，我们需要不同的数据，接下来分别从各个模块分析，每个模块需要的数据。用户基本信息就是用户的浏览行为信息分析，也就是我们只需要pageview事件就可以了；浏览器信息分析以及地域信息分析其实就是在用户基本信息分析的基础上添加浏览器和地域这个维度信息，其中浏览器信息我们可以通过浏览器的window.navigator.userAgent来进行分析，地域信息可以通过nginx服务器来收集用户的ip地址来进行分析，也就是说pageview事件也可以满足这两个模块的分析。外链数据分析以及用户浏览深度分析我们可以在pageview事件中添加访问页面的当前url和前一个页面的url来进行处理分析，也就是说pageview事件也可以满足这两个模块的分析。订单信息分析要求pc端发送一个订单产生的事件，那么对应这个模块的分析，我们需要一个新的事件chargeRequest。对于事件分析我们也需要一个pc端发送一个新的事件数据，我们可以定义为event。除此之外，我们还需要设置一个launch事件来记录新用户的访问。

|  |  |
| --- | --- |
| **最终分析模块** | **PC端js sdk事件** |
| 用户基本信息分析 | pageview事件 |
| 浏览器信息分析 |
| 地域信息分析 |
| 外链数据分析 |
| 用户浏览深度分析 |
| 订单信息分析 | chargeRequest事件 |
| 事件分析 | event事件 |
|  | launch事件 |

**Launch事件**：当用户第一次访问网站的时候触发该事件，不提供对外调用的接口，只实现该事件的数据收集。

|  |  |
| --- | --- |
| 方法名称 |  |
| 发送的数据 | u\_sd=8E9559B3-DA35-44E1-AC98-85EB37D1F263&c\_time=1449137597974&ver=1&en=e\_l&pl=website&sdk=js&b\_rst=1920\*1080&u\_ud=12BF4079-223E-4A57-AC60-C1A04D8F7A2F&b\_iev=Mozilla%2F5.0%20(Windows%20NT%206.1%3B%20WOW64)%20AppleWebKit%2F537.1%20(KHTML%2C%20like%20Gecko)%20Chrome%2F21.0.1180.77%20Safari%2F537.1&l=zh-CN |

**Pageview事件：**当用户访问页面/刷新页面的时候触发该事件。该事件会自动调用，也可以让程序员手动调用。

|  |  |
| --- | --- |
| 方法名称 | onPageView |
| 发送的数据 | ver=1&en=e\_pv&pl=website&sdk=js&b\_rst=1920\*1080&u\_ud=12BF4079-223E-4A57-AC60-C1A04D8F7A2F&b\_iev=Mozilla%2F5.0%20(Windows%20NT%206.1%3B%20WOW64)%20AppleWebKit%2F537.1%20(KHTML%2C%20like%20Gecko)%20Chrome%2F21.0.1180.77%20Safari%2F537.1&l=zh-CN&u\_sd=8E9559B3-DA35-44E1-AC98-85EB37D1F263&c\_time=1449137597979&ht=www.bjsxt.com%3A8080&p\_url=http%3A%2F%2Fwww.bjsxt.com%3A8080%2Fvst\_track%2Findex.html |

**chargeRequest事件：**当用户下订单的时候触发该事件，该事件需要程序主动调用。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方法名称 | | | onChargeReque | | |
| 发送的数据 | u\_sd=8E9559B3-DA35-44E1-AC98-85EB37D1F263&c\_time=1449139048231&  oid=orderid123&on=%E4%BA%A7%E5%93%81%E5%90%8D%E7%A7%B0&  cua=1000&cut=%E4%BA%BA%E6%B0%91%E5%B8%81&pt=%E6%B7%98%E5%AE%9D&  ver=1&en=e\_crt&pl=website&sdk=js&b\_rst=1920\*1080&  u\_ud=12BF4079-223E-4A57-AC60-C1A04D8F7A2F&b\_iev=Mozilla%2F5.0%20  (Windows%20NT%206.1%3B%20WOW64)%20AppleWebKit%2F5  37.1%20(KHTML%2C%20like%20Gecko)%20Chrome%2F21.0.  1180.77%20Safari%2F537.1&l=zh-CN | | | | |
| 参数 | | 类型 | | 是否必填 | 描述 |
| orderId | | string | | 是 | 订单id |
| orderName | | String | | 是 | 产品购买描述名称 |
| currencyAmount | | double | | 是 | 订单价格 |
| currencyType | | String | | 是 | 货币类型 |
| paymentType | | String | | 是 | 支付方式 |

**Event事件:**

当访客/用户触发业务定义的事件后，前端程序调用该方法。

|  |  |  |  |
| --- | --- | --- | --- |
| 方法名称 | onEventDuratie | | |
| 发送的数据 | ca=%E7%B1%BB%E5%9E%8B&ac=%E5%8A%A8%E4%BD%9C&c\_time=144913951266  5&u\_sd=8E9559B3-DA35-44E1-AC98-85EB37D1F263&kv\_p\_url=http%3A%2F%2Fwwwbjsxt..com%3A8080%  2Fvst\_track%2Findex.html&kv\_%E5%B1%9E%E6%80%A7key=%E5%B1  %9E%E6%80%A7value&du=1000&ver=1&en=e\_e&pl=website&sdk=js&  b\_rst=1920\*1080&u\_ud=12BF4079-223E-4A57-AC60-C1A04D8F7A2F&b\_iev=Mozilla%2F5.0%20(Windows%20NT%206.  1%3B%20WOW64)%20AppleWebKit%2F537.1%20(KHTML%2C%20lik  e%20Gecko)%20Chrome%2F21.0.1180.77%20Safari%2F537.1&l=zh-CN | | |
| 参数 | 类型 | 是否必填 | 描述 |
| category | string | 是 | 自定义事件名称 |
| action | String | 是 | 自定义事件动作 |
| map | map | 否 | 其他参数 |
| duration | long | 否 | 事件持续时间 |

**其他api方法:**在这里只介绍设置会员id的方法，其他的辅助方法到时候编写js的时候

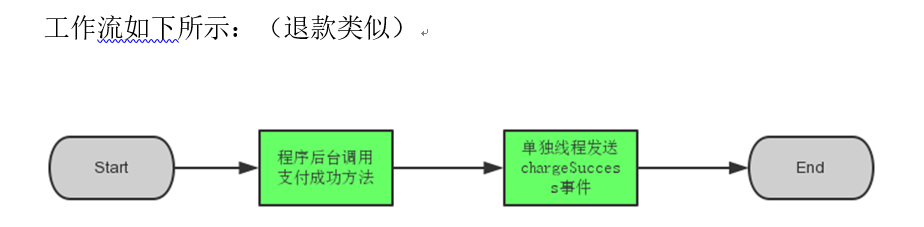
**设置会员id**

|  |  |  |  |
| --- | --- | --- | --- |
| 方法名称 | | setMemberId | |
| 方法描述 | | 该方法的主要作用是设置会员id，当会员登录后，请调用该方法进行会员id的设置。 | |
| 参数 | 类型 | 是否必填 | 描述 |
| memberid | string | 是 | 会员id |

**数据参数说明:**在各个不同事件中收集不同的数据发送到nginx服务器，但是实际上这些收集到的数据还是有一些共性的。下面将所用可能用到的参数描述如下：

|  |  |  |
| --- | --- | --- |
| 参数名称 | 类型 | 描述 |
| en | string | 事件名称, eg: e\_pv |
| ver | string | 版本号, eg: 0.0.1 |
| pl | string | 平台, eg: website |
| sdk | string | Sdk类型, eg: js |
| b\_rst | string | 浏览器分辨率，eg: 1800\*678 |
| b\_iev | string | 浏览器信息useragent |
| u\_ud | string | 用户/访客唯一标识符 |
| l | string | 客户端语言 |
| u\_mid | string | 会员id，和业务系统一致 |
| u\_sd | string | 会话id |
| c\_time | string | 客户端时间 |
| p\_url | string | 当前页面的url |
| p\_ref | string | 上一个页面的url |
| tt | string | 当前页面的标题 |
| ca | string | Event事件的Category名称 |
| ac | string | Event事件的action名称 |
| kv\_\* | string | Event事件的自定义属性 |
| du | string | Event事件的持续时间 |
| oid | string | 订单id |
| on | string | 订单名称 |
| cua | string | 支付金额 |
| cut | string | 支付货币类型 |
| pt | string | 支付方式 |

1. **Server 数据采集**



**程序后台事件分析：**本项目中在程序后台只会出发chargeSuccess事件，本事件的主要作用是发送订单成功的信息给nginx服务器。发送格式同pc端发送方式， 也是访问同一个url来进行数据的传输。格式为:

|  |  |
| --- | --- |
| 最终分析模块 | PC端js sdk事件 |
| 订单信息分析 | chargeSuccess事件  chargeRefund事件 |

chargeSuccess事件：当会员最终支付成功的时候触发该事件，该事件需要程序主动调用。

|  |  |  |  |
| --- | --- | --- | --- |
| 方法名称 | | onChargeSuccess | |
| 发送的数据 | | u\_mid=bjsxt&c\_time=1449142044528&oid=orderid123&ver=1&en=e\_cs&pl=javaserver&sdk=jdk | |
| 参数 | 类型 | 是否必填 | 描述 |
| orderId | string | 是 | 订单id |
| memberId | string | 是 | 会员id |

chargeRefund事件：当会员进行退款操作的时候触发该事件，该事件需要程序主动调用。

|  |  |  |  |
| --- | --- | --- | --- |
| 方法名称 | | onChargeRefund | |
| 发送的数据 | | u\_mid=bjsxt&c\_time=1449142044528&oid=orderid123&ver=1&en=e\_cr&pl=jdk&sdk=java | |
| 参数 | 类型 | 是否必填 | 描述 |
| orderId | string | 是 | 订单id |
| memberId | string | 是 | 会员id |

集成方式：直接将java的sdk引入到项目中即可，或者添加到classpath中。

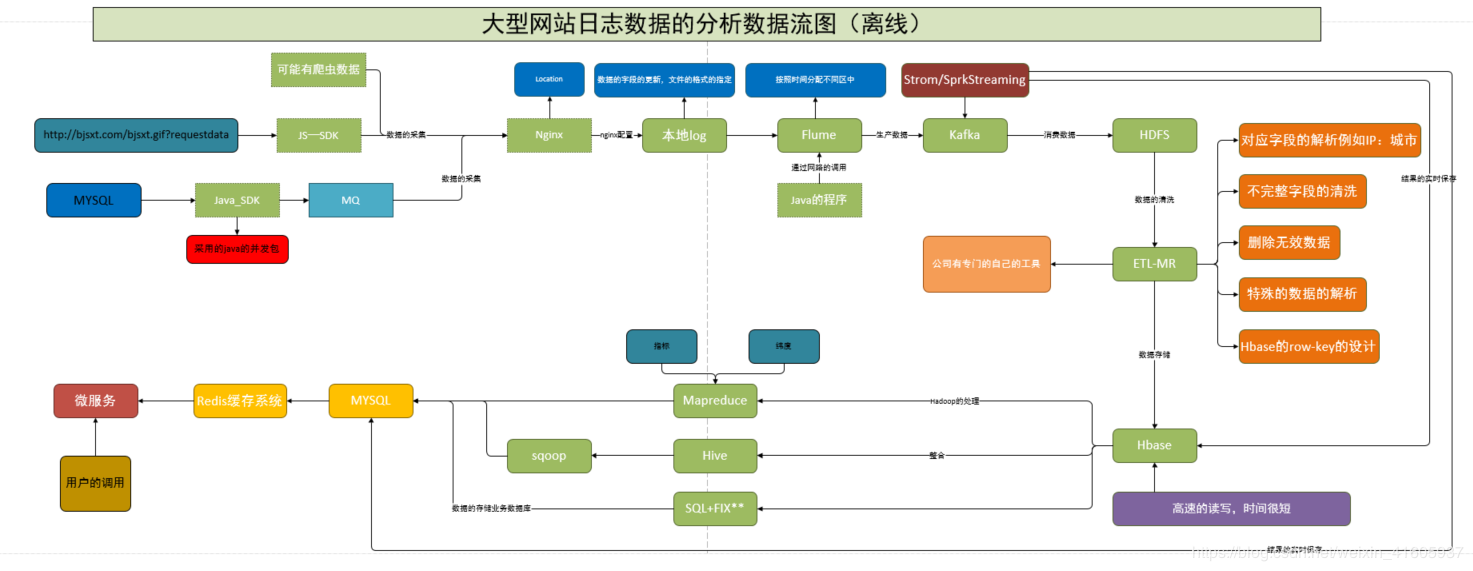
数据参数说明参数描述如下：

|  |  |  |
| --- | --- | --- |
| 参数名称 | 类型 | 描述 |
| en | string | 事件名称, eg: e\_cs |
| ver | string | 版本号, eg: 0.0.1 |
| pl | string | 平台, eg: website,javaweb,php |
| sdk | string | Sdk类型, eg: java |
| u\_mid | string | 会员id，和业务系统一致 |
| c\_time | string | 客户端时间 |
| oid | string | 订单id |

[**项目数据清洗**](https://blog.csdn.net/weixin_41605937/article/details/107507048#%E9%A1%B9%E7%9B%AE%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97)

[**项目的业务**](https://blog.csdn.net/weixin_41605937/article/details/107507048#%E9%A1%B9%E7%9B%AE%E7%9A%84%E4%B8%9A%E5%8A%A1)

[**项目系统优化**](https://blog.csdn.net/weixin_41605937/article/details/107507048#%E9%A1%B9%E7%9B%AE%E7%B3%BB%E7%BB%9F%E4%BC%98%E5%8C%96)

​​​

**数据处理流程五大步骤**

1. 数据采集  
   Flume： 使用Flume对数据进行采集，将web日志写入到HDFS
2. 数据清洗  
   使用Spark、Hive、MapReduce或者其他的一些分布式计算框架  
   清洗完之后的数据可以存放在HDFS或者Hive、Spark SQL里
3. 数据处理  
   按照我们的需要进行相应业务的统计和分析（使用Spark、Hive、MapReduce、Flink等框架）
4. 数据处理结果入库  
   结果可以存放到RDBMS、NoSQL等数据库
5. 数据的可视化  
   通过图形化展示的方式展现出来：饼图、柱状图、地图、折线图  
   工具：ECharts、HUE、Zeppelin、Kibana等