|  |  |
| --- | --- |
| 密级： | 公开 |
| 版本： | V1.0 |



第一创业BDI实时统计后台详细设计

2017年03月22日

**版权声明**

本使用手册版权由百分点信息科技有限公司所有，未经百分点信息科技有限公司的许可和授权，任何机构及人员不得将本手册及本使用手册包含的任何文字、内容及图片以任何理由，任何方式或手段（电子的或机械的）进行复制或传播。

**修订记录**

|  |  |  |  |
| --- | --- | --- | --- |
| 版 本 | 修订说明 | 编制/日期 | 审核/日期 |
| V1.0 | 第一创业BDI实时统计后台详细设计 | 刘剑/  2017.03.22 |  |
|  |  |  |  |
|  |  |  |  |

目 录

[一、 序言 1](#_Toc477945515)

[1.1目的 1](#_Toc477945516)

[二、 总体设计 1](#_Toc477945517)

[2.1系统运行环境 1](#_Toc477945518)

[2.2系统总体目标 1](#_Toc477945519)

[三、 后端详细设计 2](#_Toc477945520)

[3.1、系统整体架构 2](#_Toc477945521)

[3.2 服务设计 3](#_Toc477945522)

[3.2.1整体设计 3](#_Toc477945523)

[3.2.2定时topic切换任务逻辑 4](#_Toc477945524)

[3.1.3 sparkstreaming逻辑 5](#_Toc477945525)

[四、 后台配置 5](#_Toc477945526)

[4.1系统配置说明 6](#_Toc477945527)

[五、 系统缺陷 6](#_Toc477945528)

# 序言

## 1.1目的

记录第一创业BDI实时统计设计，供后续人员使用，特撰写了本文档。

# 总体设计

## 2.1系统运行环境

|  |  |
| --- | --- |
| 组件 | 版本 |
| 星环平台 | 4.6.2 |

## 2.2系统总体目标



实时访客：

每分钟内的pv，uv，ip数，每个session的明细数据。

页面医生：

每个页面的pv。

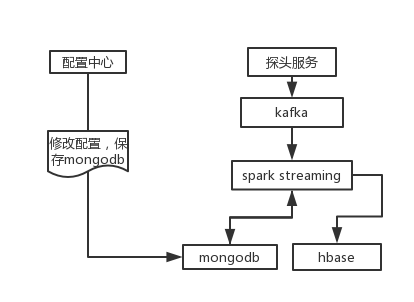
每个页面每个元素的总点击数。

每个页面每个元素分小时的点击数。

每个页面每个元素分小时的点击来源。

# 后端详细设计

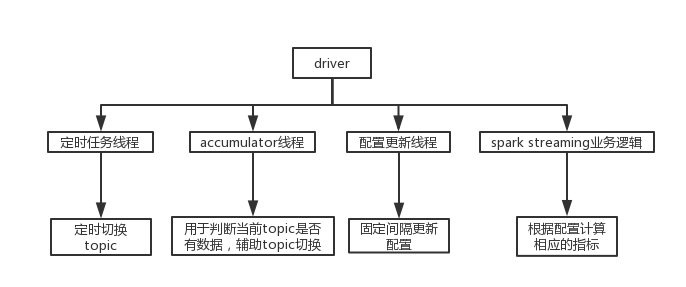
## 3.1、系统整体架构



配置中心用于运营，增减关注的对象。

## 3.2 服务设计

### 3.2.1整体设计

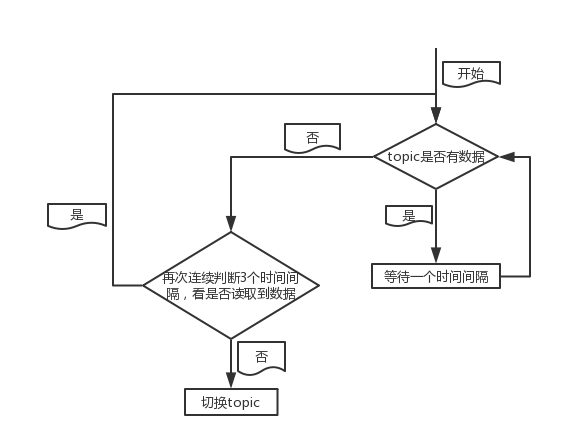


由于kafka数据每天都有一个topic，为了使spark streaming自动切换topic，增加了两个后台线程，accumulator线程用于计数，辅助定时任务线程决定是否马上切换，防止当前topic数据没有处理完全。

为了使运营配置在不重启应用的情况下生效，增加了配置更新线程，线程每隔固定间隔生效最新配置。

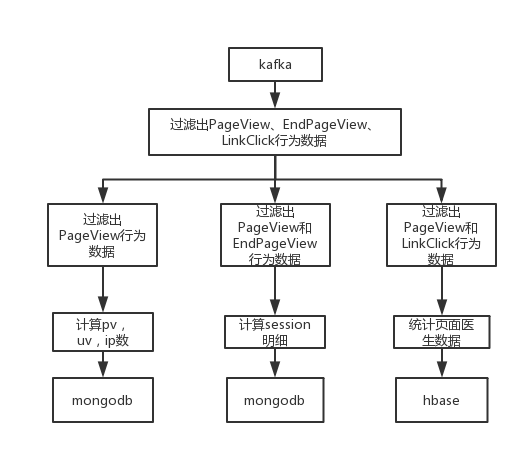
spark streaming是核心逻辑，用于计算上述指标。

### 3.2.2定时topic切换任务逻辑



topic切换线程，作用是保证在当前数据源无数据的情况下切换数据源。

### sparkstreaming逻辑



pv指标计算：对于每一个PageView行为计数。

uv和ip数计算：对于每一条PageView行为，取gid计算uv，取ip计算独立ip数。

session明细计算：取一个session内的数据做合并。

页面医生计算：分多个维度进行计数。

# 后台配置

配置均采用java的properties文件格式。

## 4.1系统配置说明

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
| kafka.zk.address | kafka使用的zookeeper路径 |
| kafka.group | 消费组 |
| interval.second | spark streaming batch间隔 |
| num.kafka.read.thread | 读取kafka的线程数 |
| kafka.data.hasprefix | 数据是{}{}还是{} |
| topic.switch.hour | topic切换的小时点 |
| topic.switch.minute | topic切换的分钟点 |
| hbase.rootdir | hbase在hdfs的根目录 |
| hbase.zookeeper.quorum | hbase使用的zookeeper集群 |
| hbase.zookeeper.property.clientPort | hbase使用的zookeeper端口 |
| zookeeper.znode.parent | hbase在zookeeper的路径 |
| hbase.table | 页面医生数据存储的hbase表 |
| mongo.uri | mongodb连接url |
| mongo.db | 数据使用的mongodb库 |
| mongodb.appkey.table | 存储appkey的collection |
| mongodb.reg.table | 存储正则的collection |
| mongodb.doctor.table | 存储页面医生需要统计的url |
| mongodb.stat.table | 实时统计pv、uv、ip存储collection |
| mongodb.detail.table | 实时统计session明细存储collection |
| mongodb.searchengine.table | 搜索引擎配置collection |
| detial.ip.location.method | ip转换规则 |
| card.init.bits | 基数估计使用的bit数 |
| conf.update.interval.minute | 配置更新的间隔 |

# 系统缺陷

系统和mongodb交互次数频繁，当数据量较大时，系统瓶颈依赖于mongodb的写入和读取速度。

# 服务启动

nohup spark-submit --class com.bfd.spark.bdi\_realtime.BDIRealtime --master yarn --name bdi\_realtime --conf spark.streaming.backpressure.enabled=true --conf spark.streaming.receiver.maxRate=200000 --files log4j.properties,bdi\_realtime.properties --num-executors 10 --driver-memory 10g --executor-memory 9g --executor-cores 20 bdi\_realtime-1.0.0-jar-with-dependencies.jar bdi\_realtime.properties log4j.properties &