# 课程介绍

课程名称：

点击流日志分析

1、什么是点击流系统？记录用户在网站上的操作，用户行为轨迹。

2、日志有哪些需要注意的地方，如何采集日志（flume），日志格式，日志包含的信息量（字段）

3、分析什么？

网址来源，TOPK

客户端流量占比 Android、IOS、.....

网页热力图

课程目标：

1. 掌握点击流系统的架构及工作原理
2. 掌握点击点击流中常见的字段及其业务含义
3. 掌握点击流分析系统开发

课程大纲：

1. 背景知识
2. 需求分析
3. 架构设计
4. Storm程序开发
5. 同步程序开发

# 课程内容

# 1、背景知识

## 1.1、什么是流量日志

专业名词：点击流，一般叫做点击流日志分析

网站分析的主要手段是分析网站的点击流数据

点击流这个概念更注重用户浏览网站的整个流程，我们一般也叫做用户行为轨迹

总结：点击流其实就是用户日常浏览你的网站时产生的日志信息

## 1.2、如何确定用户的行为轨迹数据？

在一个**sessinon** 会话中的操作行为（浏览行为，点击行为），每个操作都会产生一条日志信息

## 1.3、点击流日志信息中包含什么东西？

1、来源URL ref

2、当前URL req http://www.itcast.cn/121212.html

3、请求时间 reqTime （用户打开网站的时间）

4、操作类型 type （0：浏览操作，1：点击操作）

9、页面停留时间

10、用户IP地址

11、sessionID

12、用户的账号信息

13、点击什么东西

5、操作系统

6、浏览器的信息

7、屏幕尺寸

8、鼠标点击的位置

## 1.4、网站分析基础指标

* IP(独立IP)： 即Internet　Protocol,指独立IP数。00:00-24:00内相同IP地址之被计算一次。
* PV(访问量)： 即Page View,即页面浏览量或点击量，用户每次刷新即被计算一次。
* UV(独立访客)：即Unique　Visitor,访问您网站的一台电脑客户端为一个访客。

00:00-24:00内相同的客户端只被计算一次。

**用户访问量>=独立访客>=IP**

* Source: 用户来源域名的统计
* Browser: 用户的访问设备统计

## 网站分析指标分类

### 1.5.1流量数量指标

#### 1.5.1.1浏览量（PV）

* 定义：页面浏览量即为PV(Page View)，用户每打开一个页面就被记录 1 次。
* 技术说明：一个PV即电脑从网站下载一个页面的一次请求。当页面上的JS文件加载后，统计系统才会统计到这个页面的浏览行为，有如下情况需注意：

1.用户多次打开同一页面，浏览量值累计。

2.如果客户端已经有该缓冲的文档，甚至无论是不是真的有这个页面（比如JavaScript生成的一些脚本功能），都可能记录为一个PV。但是如果利用网站后台日志进行分析，因为缓存页面可能直接显示而不经过服务器请求，那么不会记录为一个 PV。

* 涵义：PV越多越说明该页面被浏览的越多。PV之于网站，就像收视率之于电视，已成为评估网站表现的基本尺度。

#### 1.5.1.2访问次数

* 定义：访问次数即Visit，访客在网站上的会话(Session)次数，一次会话过程中可能浏览多个页面。
* 技术说明：如果访客连续30分钟内没有重新打开和刷新网站的网页，或者访客关闭了浏览器，则当访客下次访问您的网站时，访问次数加1。反之，访客离开后半小时内再返回，则算同一个访次，以上对访客的判断均以 Cookie 为准。
* 涵义：页面浏览量（PV）是以页面角度衡量加载次数的统计指标，而访问次数（Visit）则是访客角度衡量访问的分析指标。如果网站的用户黏性足够好，同一用户一天中多次登录网站，那么访问次数就会明显大于访客数。

#### 1.5.1.3访客数（UV）

* 定义：访客数（UV）即唯一访客数，一天之内网站的独立访客数(以Cookie 为依据 )，一天内同一访客多次访问网站只计算 1 个访客。
* 技术说明：当客户端第一次访问某个网站服务器的时候，网站服务器会给这个客户端的电脑发一个Cookie，记录访问服务器的信息。当下一次再访问服务器的时候，服务器就可以直接找到上一次它放进去的这个Cookie，如果一段时间内，服务器发现两个访次对应的Cookie编号一样，那么这些访次一定就是来自一个 UV 了。
* 涵义：唯一访客数（UV）是访客维度看访客到达网站的数量。

#### 1.5.1.4新访客数

* 定义：一天的独立访客中，历史第一次访问网站的访客数。
* 涵义：新访客数可以衡量营销活动开发新用户的效果。

#### 1.5.1.5新访客比率

* 定义：新访客比率=新访客数/访客数。即一天中新访客数占总访客数的比例。
* 涵义：整体访客数不断增加，并且其中的新访客比例较高，能表现网站运营在不断进步。就像人体的血液循环一样，有新鲜的血液不断补充进来，充满活力。

#### 1.5.1.6 IP数

* 定义：一天之内，访问网站的不同独立IP个数加和。其中同一IP无论访问了几个页面，独立 IP 数均为 1。
* 涵义：从 IP 数的角度衡量网站的流量。

### 1.5.2流量质量指标

#### 1.5.2.1跳出率

* 定义：只浏览了一个页面便离开了网站的访问次数占总的访问次数的百分比，即只浏览了一个页面的访问次数 / 全部的访问次数汇总。
* 涵义：跳出率是非常重要的访客黏性指标，它显示了访客对网站的兴趣程度：跳出率越低说明流量质量越好，访客对网站的内容越感兴趣，这些访客越可能是网站的有效用户、忠实用户。该指标也可以衡量网络营销的效果，指出有多少访客被网络营销吸引到宣传产品页或网站上之后，又流失掉了，可以说就是煮熟的鸭子飞了。比如，网站在某媒体上打广告推广，分析从这个推广来源进入的访客指标，其跳出率可以反映出选择这个媒体是否合适，广告语的撰写是否优秀，以及网站入口页的设计是否用户体验良好。

#### 1.5.2.2平均访问时长

* 定义：平均每次访问在网站上的停留时长，即平均访问时长等于总访问时长与访问次数的比值。
* 涵义：平均访问时间越长则说明访客停留在网页上的时间越长：如果用户对网站的内容不感兴趣，则会较快关闭网页，那么平均访问时长就短；如果用户对网站的内容很感兴趣，在网站停留了很长时间，平均访问时长就长。

#### 1.5.2.3平均访问页数

* 定义：平均每次访问浏览的页面数量，平均访问页数=浏览量/访问次数。
* 涵义：平均访问页数多说明访客对网站兴趣越大。而浏览信息多也使得访客对网站更加了解，这对网站市场信息的传递，品牌印象的生成，以至于将来的销售促进都是有好处的。一般来说，会将平均访问页数和平均访问时长这两个指标放在一起分析，进而衡量网站的用户体验情况。

### 1.5.3流量转化指标

#### 1.5.3.1转化次数

* 定义：访客到达转化目标页面，或完成网站运营者期望其完成动作的次数。
* 涵义：转化就是访客做了任意一项网站管理者希望访客做的事。与网站运营者期望达到的推广目的和效果有关。

#### 1.5.3.2转化率

* 定义：转化率 = 转化次数 / 访问次数。
* 涵义：转化率即访问转化的效率，数值越高说明越多的访次完成了网站运营者希望访客进行的操作。

用户在电商的行为：

用户 浏览某一个专题页、流量加入购物车、下订单、支付

a 专题页1 1 1 1

b 专题页1 1 0 0

c 专题页1 1 0 0

d 专题页1 1 1 0

e 专题页1 1 1 0

前提是有浏览过：

A转化： 3/5 = 60% 加入购物到下单的转化

B 转换： 1/3 = 33% 下单到支付的转化

支付的问题：支付密码太长，短信没有收到，支持的银行太少，不能货到付款，浏览器的兼容性

重复次数太多，30% 因为密码多次输入造成订单流失，提交点击行为获取

短信没收到，10% 因为短信验证码输入框没有输入，确定记录没有

支持的银行太少，5%，因支付银行列表中没有找到对应的银行

## 1.6 企业点击流日志分析的一般流程

1、数据准备工作

有点击流日志系统开发人员，将网站的日志信息全部收集过来。日志信息中包含的指标尽可能多

2、业务需求(每个部门都有很多数据分析的需求，比如专题页就有很多个)

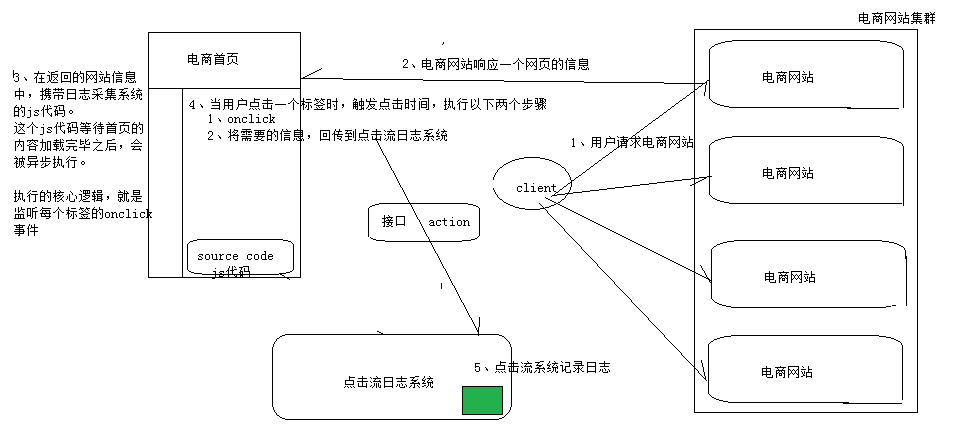
需求人员需要对指标进行标准定义，根据指标列出要收集哪些数据

3、报表开发

报表开发人员评估业务需求中的指标，如果指标需要的数据没有，进行反馈。

巧妇难做无米之炊

## 1.7、点击流系统的一般工作原理



# 2、需求分析

## 2.1、点击流日志的主要分类

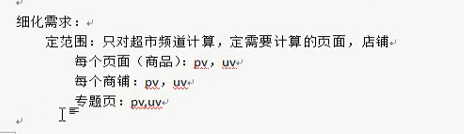
* 搜索日志：搜索历史的记录主要包括用户信息、时间、地址、检索的关键词，检索关键词拼音及缩写，用户年纪等，其中，记录的时间包括检索发生时的小时、当天是周几、当天的日期信息；地址信息包括了省市区县信息。
* 点击日志：点击历史记录了当前点击记录的用户信息、时间、地址、检索词、点击记录的序号、点击记录的ID，其中，记录的时间包括检索发生时的小时、当天是周几、当天的日期信息；记录了该产品是在搜索结果中点击的还是推荐结果中点击的；地址信息包括了省市区县信息。
* 购买日志：购买历史分类已付款和未付款，并记录的有付款时间和订单时间。统计还记录了用户信息、时间、地址、检索词、点击记录的序号、购买记录的ID，并且统计了在查看了该记录多少次以后购买的，也记录了该产品是在搜索结果中点击的还是推荐结果中点击的，其中，记录的时间包括检索发生时的小时、当天是周几、当天的日期信息；地址信息包括了省市区县信息。
* 浏览日志：浏览数据是用户在查看产品信息的浏览记录，一次点击查看的页面会记录多条浏览数据，该记录是采样数据，采集的频率是10秒一次，记录了用户信息、时间、地址、检索词、产品ID、当前鼠标浏览位置、当前页面焦点位置、当前页面滚动次数、距离上一次滚动时间等信息。

## 2.2、点击流日志的采集方式

* 显式模式主要是面谈和样本调查，用户在被调查过程中会受到心理等外界因素影响，有一定的局限性，并且调查成本较大；
* 隐式模式主要是对用户在访问过程中产生的数据进行日志分析，相对客观且对用户行为有较小的干涉，是用户行为分析中常用方式。

## 2.3、业务需求

  在电商企业中，产品在做活动的时候，需要根据一些指标数据来观看活动的效果，常见的会包括pv、uv、订单数、订单人数、订单转化率、用户来源等等指标。随着业务人员对数据的实时性的要求越来越高，一些不依赖历史数据、可以被快速分析的指标就会使用流式计算来分析，让业务人员实时的看到活动推动的效果，并辅助做相关的决策。

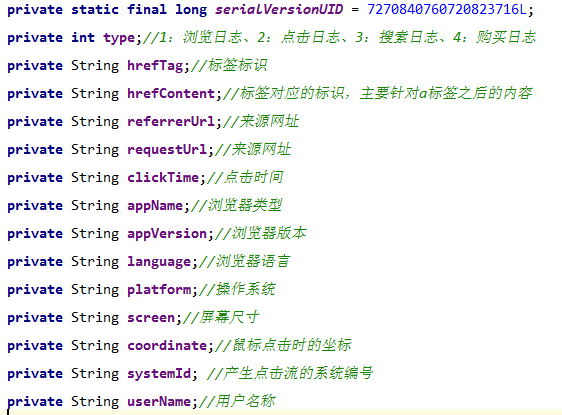


## 2.4、功能需求

1. 实时计算一个页面的访问量、独立访客、独立IP、访问次数
2. 实时计算一个标签的访问量、独立访客、独立IP、访问次数
3. 对相关实时指标通过曲线图进行展示、维度为每分钟、每半小时、每小时、每天等

## 2.5、数据准备

通过点击流系统获取用户行为日志，经过解析之后，得到以下的字段



## 2.6、数据分析

* type：表示日志类型有哪些，分别是浏览日志、点击日志、搜索日志、购买日志，通过type能够明确的知道用户当前操作的业务类型
* hrefTag：一个标签的唯一标识，通过hrefTag能够知道用户当前点击的标签是什么
* hrefContent：当前标签的文本内容
* referrefUrl：当前请求的来源网址，如：<http://www.itcast.cn/>
* requestUrl：当前请求的网址，如：<https://www.taobao.com/market/nvxie/citiao/index.php?spm=a21bo.7724922.8374-1.3.1I6CoM>
* clickTime：用户点击的时间，该时间作为用户操作最准确的时间，作为后期业务分析的唯一时间指标。
* coordinate：用户操作时，鼠标的位置，可以用来分析用户的点击习惯
* systemId：产生日志的业务系统有很多，包括商品系统、订单系统、会员系统、论坛系统等等，该字段用来区分不同的业务系统。
* userName：当前用户的账户名称

## 2.7、功能实现分析

    1，实时计算一个页面的访问量、独立访客、独立IP、访问次数

        定位一个页面可以通过用户当前请求的requestUrl进行定位。

        产品人员将需要分析的页面，输入给程序。

    2，实时计算一个标签的访问量、独立访客、独立IP、访问次数

        定位一个标签可以通过hrefTag进行分析。

        产品人员将需要分析的标签，输入给程序。

    3、如何存储用户的输入信息

        将产品人员需要分析的页面或标签，抽象为一个任务。任务信息为：任务编号、任务需要比对的字段、任务的值（用户输入的值）、比较方式（String的三种比较方式，包含，相等，正则）等。

    4、如何进行曲线展示

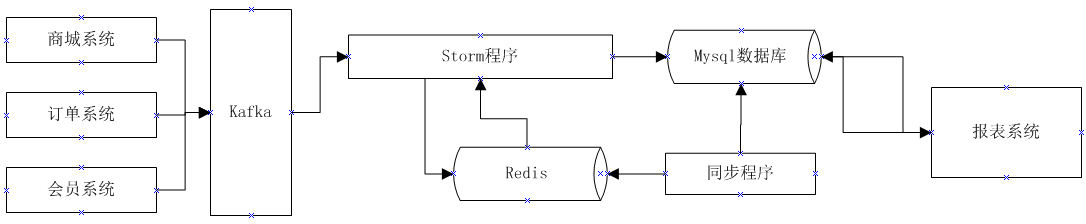
        为每个任务生成每分钟的增量数据，在增量数据的基础上进行曲线展示

# 3、原型设计

产品经理设计产品原形

# 4、架构设计

## 4.1、整体流程设计



1、通过日志收集系统将数据获取并存放到某个存储介质中，本例可以使用kafka

2、Storm程序从kafka中消费数据数据，逐条消费的日志

3、Storm程序从数据库中加载产品人员配置的任务信息

4、Storm程序计算每个任务的各项指标，各项指标的中间结果存放在Redis中

5、同步程序，按照一定的时间周期从Redis中计算每个任务的增量数据，并将增量数据存放在mysql数据库中

6、同步程序，按照每个指标的增量数据计算不同维度的基础数据

7、报表系统从Mysql数据库获取每个每个指标的基础数据进行展示

## 4.2、功能模块设计

* 数据收集模块，略
* 数据存储模块，在Kafka中创建点击流日志相关的topic
* Storm程序，该模块主要有两个个功能，

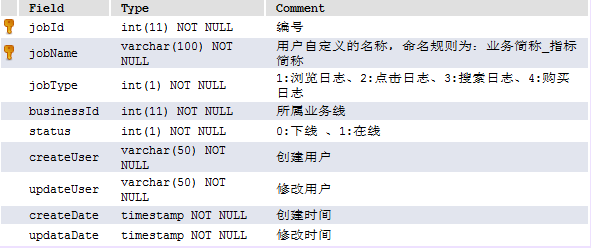
功能一：从kafka中消费数据进行实时计算，并将结果存放在Redis中

功能二：从Mysql数据库中读取配置文件并定时更新

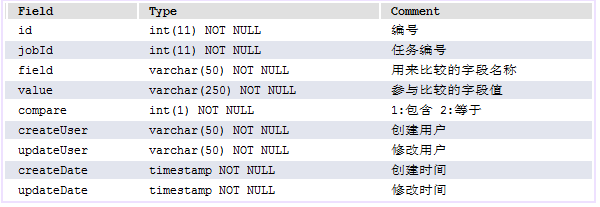
* 同步程序，该模块主要是配置相关的定时任务，主要两个方向：方向一：从Redis中读取不同任务的增量数据保存到Mysql中。方向二：根据增量数据计算不同维度的数据。
* 报表系统，从mysql数据库中读取不同任务的维度数据进行展示，略

## 数据模型设计

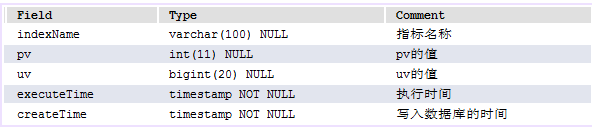
### 任务表 log\_analyze\_job



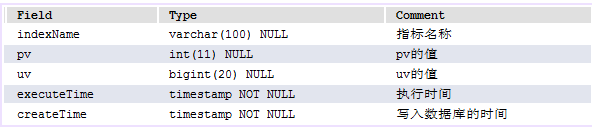
### 4.3.2、任务明细表  log\_analyze\_job\_condition



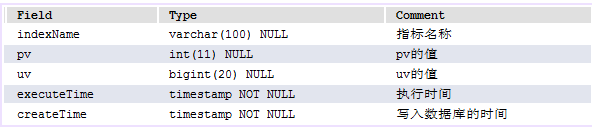
### 4.3.3、任务分钟增量数据表 log\_analyze\_job\_nimute\_append



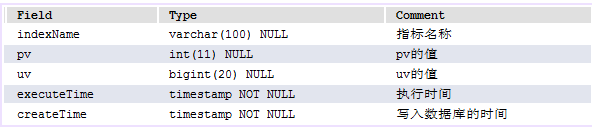
### 任务半小时增量数据表   log\_analyze\_job\_half\_append



### 任务1小时增量数据表   log\_analyze\_job\_hour\_append

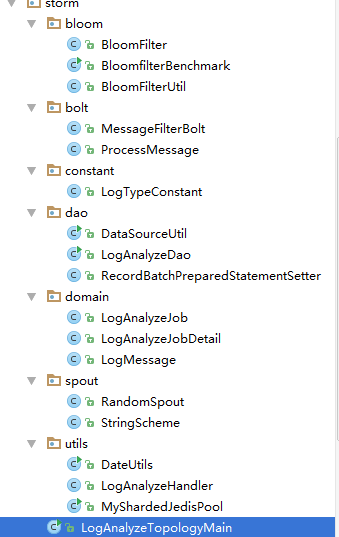


### 4.3.6、任务每天全量数据表 log\_analyze\_job\_day

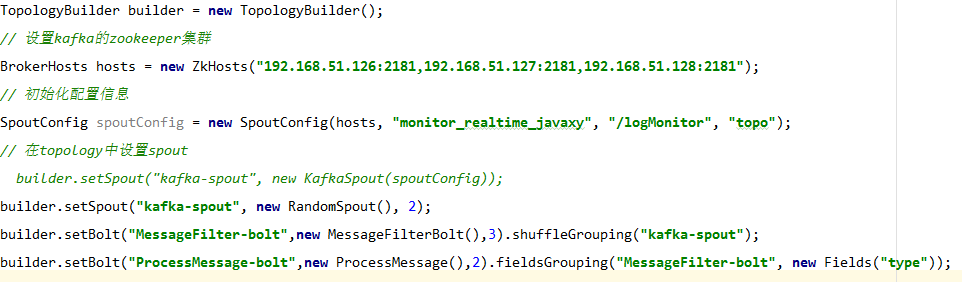


# 5、Storm程序开发

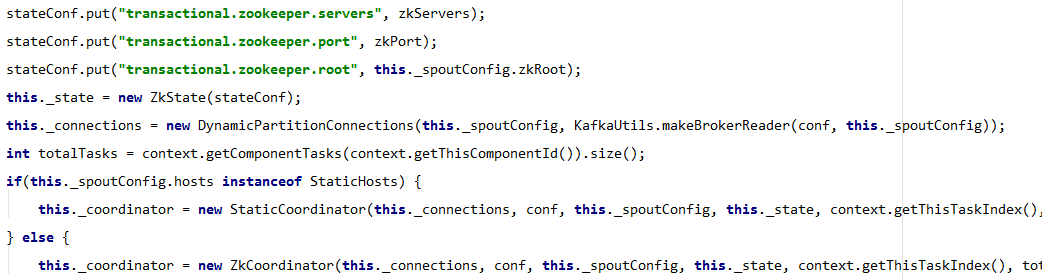
## Storm程序整体目录



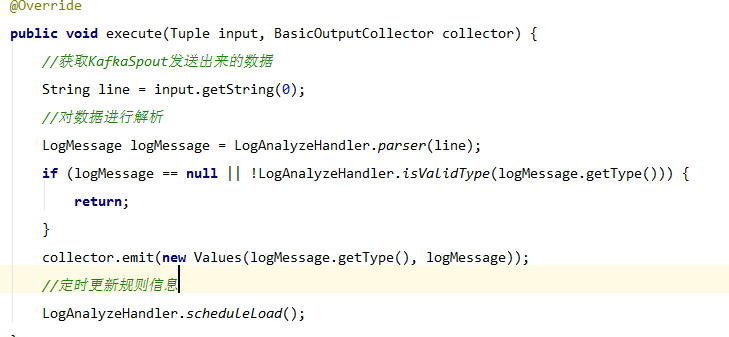
## LogAnalyzeTopologyMain 驱动类



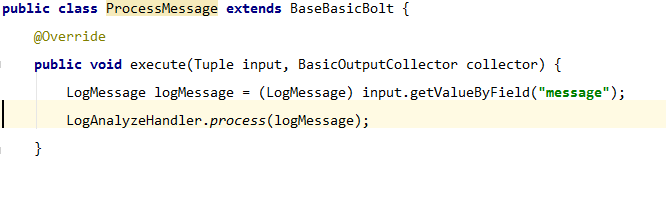
## KafkaSpout，从Kafka中读取



## MessageFilterBolt，过滤数据



## ProcessMessageBolt，计算实时指标

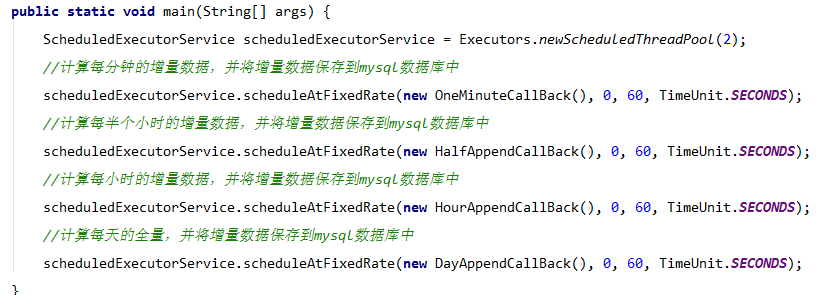


## 5.6、LogAnalyzeHandler，核心类，负责任务加载、任务计算等。

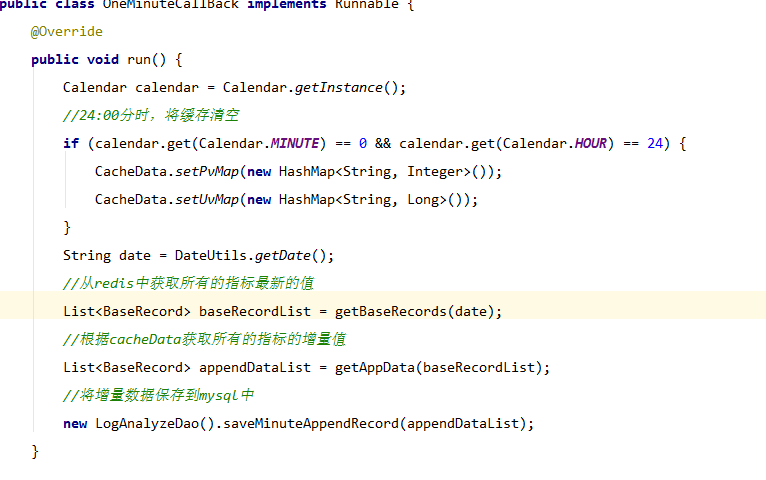


# 同步程序设计

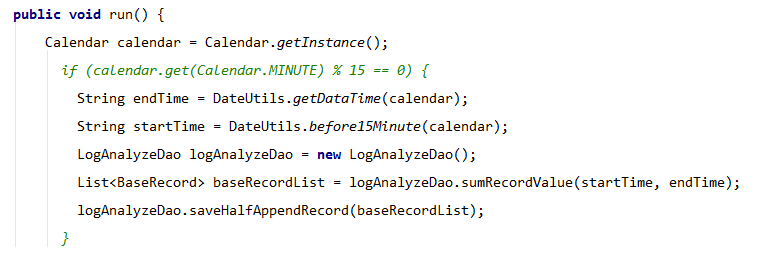
## 6.1、主函数，配置下相关的调度程序



## 6.2、OneMinuteCallBack，计算增量数据



## 6.3、HalfAppendCallBack 计算半小时的增量数据

****