

后台服务器,比如Tomcat、Jetty;但是,其实在面向 大量用户, 高并发 (每秒访问量过万) 的情况下, 通常 都不会直接是用Tomcat来接收请求。这种时候,通常, 都是用Nginx来接收请求,并且后端接入Tomcat集群/ Jetty集群,来进行高并发访问下的负载均衡。 比如说,Nginx,或者是Tomcat,你进行适当配置之 后,所有请求的数据都会作为log存储起来;接收请求的 后台系统(J2EE、PHP、Ruby On Rails),也可以按 照你的规范, 每接收一个请求, 或者每执行一个业务逻 辑,就往日志文件里面打一条log。

网站/app会发送请求到后台服务器, 通常会由Nginx接收请求,并进行转发



日志文件(通常由我们预 先设定的特殊的格式)通 常每天一份。此时呢,由 于可能有多份日志文件, 因为有多个web服务器。



网站或者app。非常重要的一点,就是埋 点。也就是说,埋点,在网站/app的哪个页 面的哪些操作发生时, 前端的代码 (网站, JavaScript; app, android/IOS), 就通过 网络请求, (Ajax; socket), 向后端的服 务器发送指定格式的日志数据。

日志文件

可以每收集1分钟的数据,或者每收集一 点数据,就放入文件,然后转移到flume 中去,或者直接通过API定制,直接把一 条一条的log打入flume。可以配置 flume, 将数据写入Kafka

实时数据, 通常都是从分布式消息队列集群中 读取的,比如Kafka;实时数据,实时的log, 实时的写入到消息队列中, 比如Kafka; 然后 呢, 再由我们后端的实时数据处理程序 (Storm、Spark Streaming), 实时从 Kafka中读取数据, log 日志。然后进行实时的 计算和处理。



实时计算

大数据实时计算系统,比如说用Storm、 Spark Streaming开发的,可以实时的从 Kafka中拉取数据,然后对实时的数据进行 处理和计算, 这里可以封装大量复杂的业务 逻辑, 甚至调用复杂的机器学习、数据挖 掘、智能推荐的算法, 然后实现实时的车辆 调度、实时推荐。

数据总线

flume负责将log日志实时传输 到Kafka中

数据存储

计算完成的数据,写入到如Mysql 等的关系型数据库中

✔ 数据可视化

计算结果进行可视化,指导PM等 优化企业运营,或者修改用户端显

②用户表 user_id、username、name、age、professional、city、sex

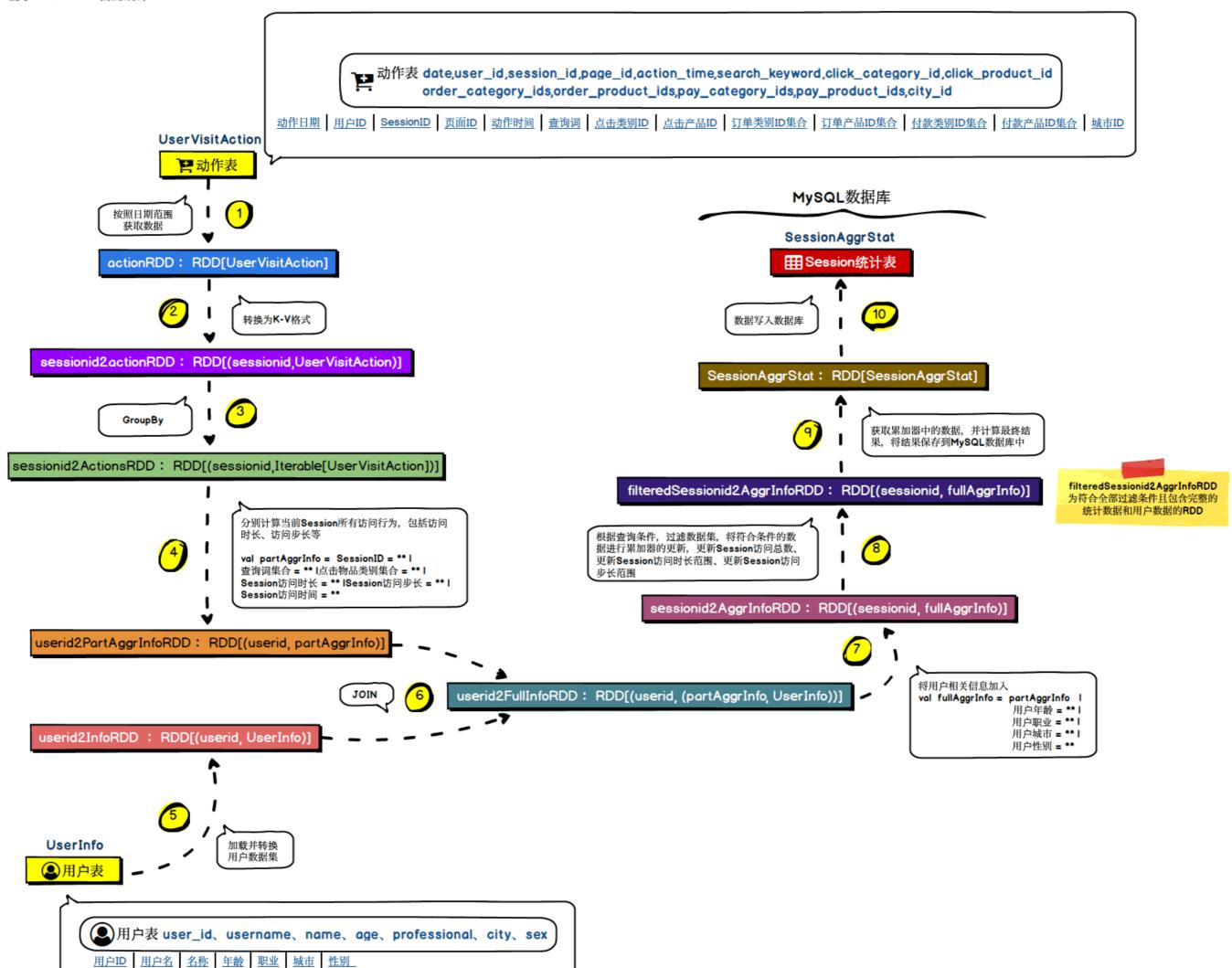
 用户ID
 用户名
 名称
 年龄
 职业
 城市
 性别

● 物品表 product_id、product_name、extend_info

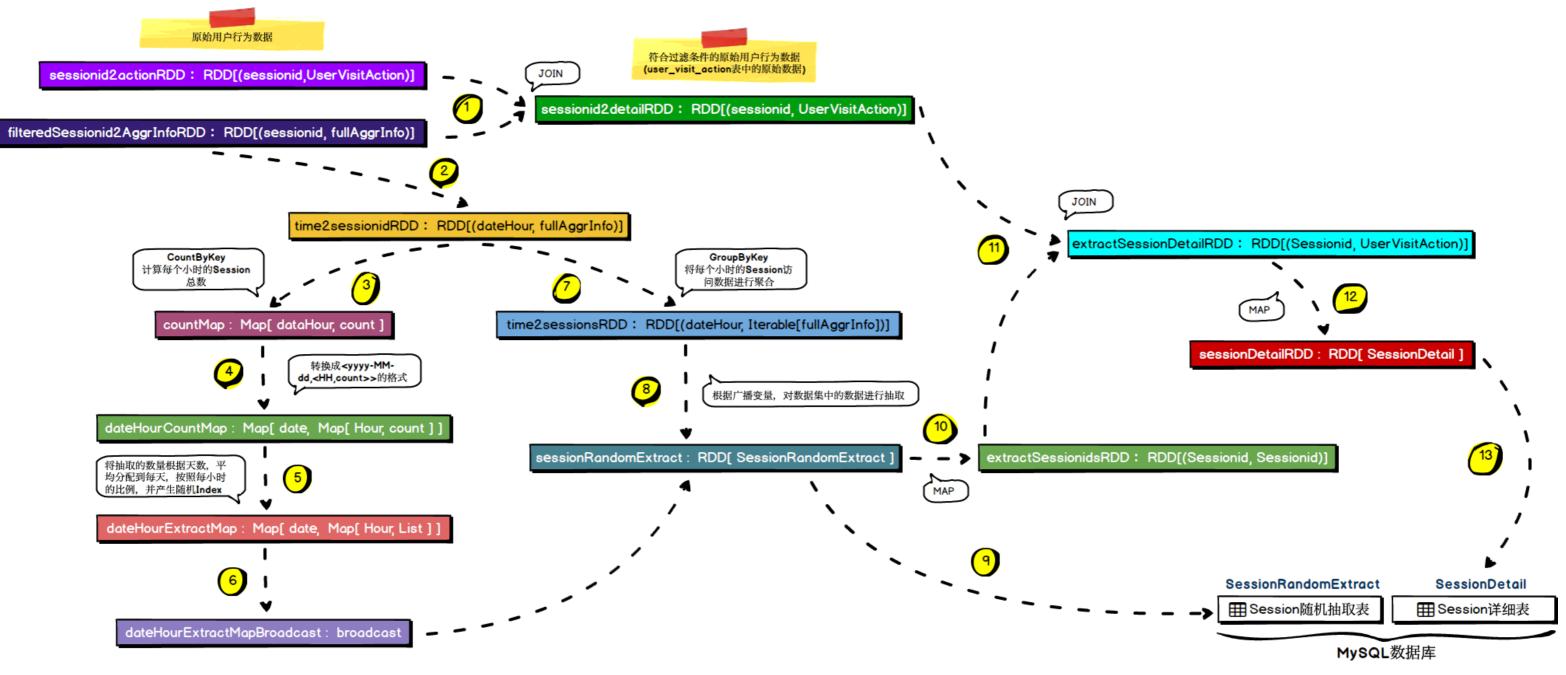
物品ID 物品名称 额外信息

动作表 date,user_id,session_id,page_id,action_time,search_keyword,click_category_id,click_product_id order_category_ids,order_product_ids,pay_category_ids,pay_product_ids,city_id

 动作日期
 用户ID
 SessionID
 页面ID
 动作时间
 查询词
 点击类别ID
 点击产品ID
 订单类别ID集合
 订单产品ID集合
 付款类别ID集合
 付款产品ID集合
 城市ID

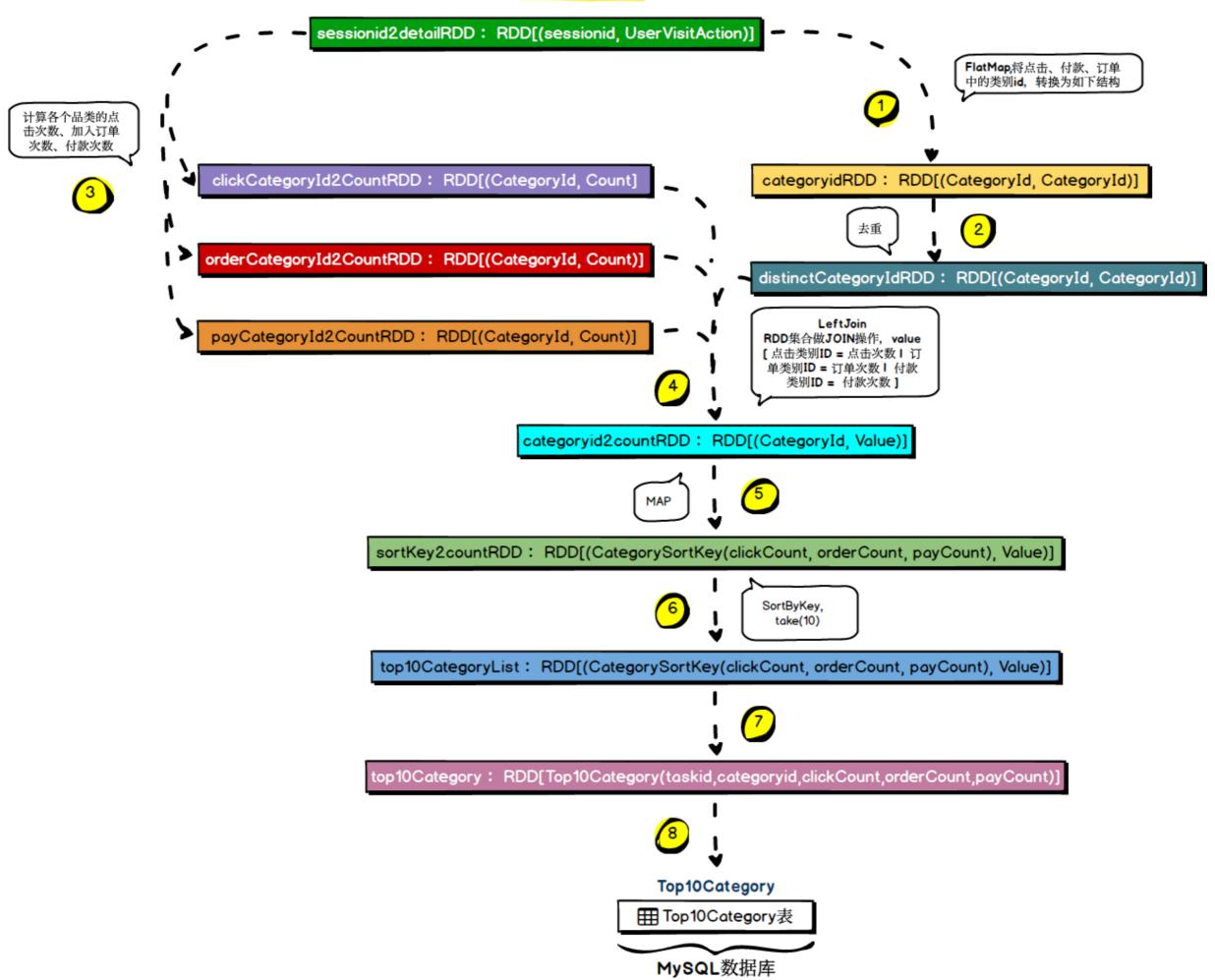


需求2: Session随机抽取



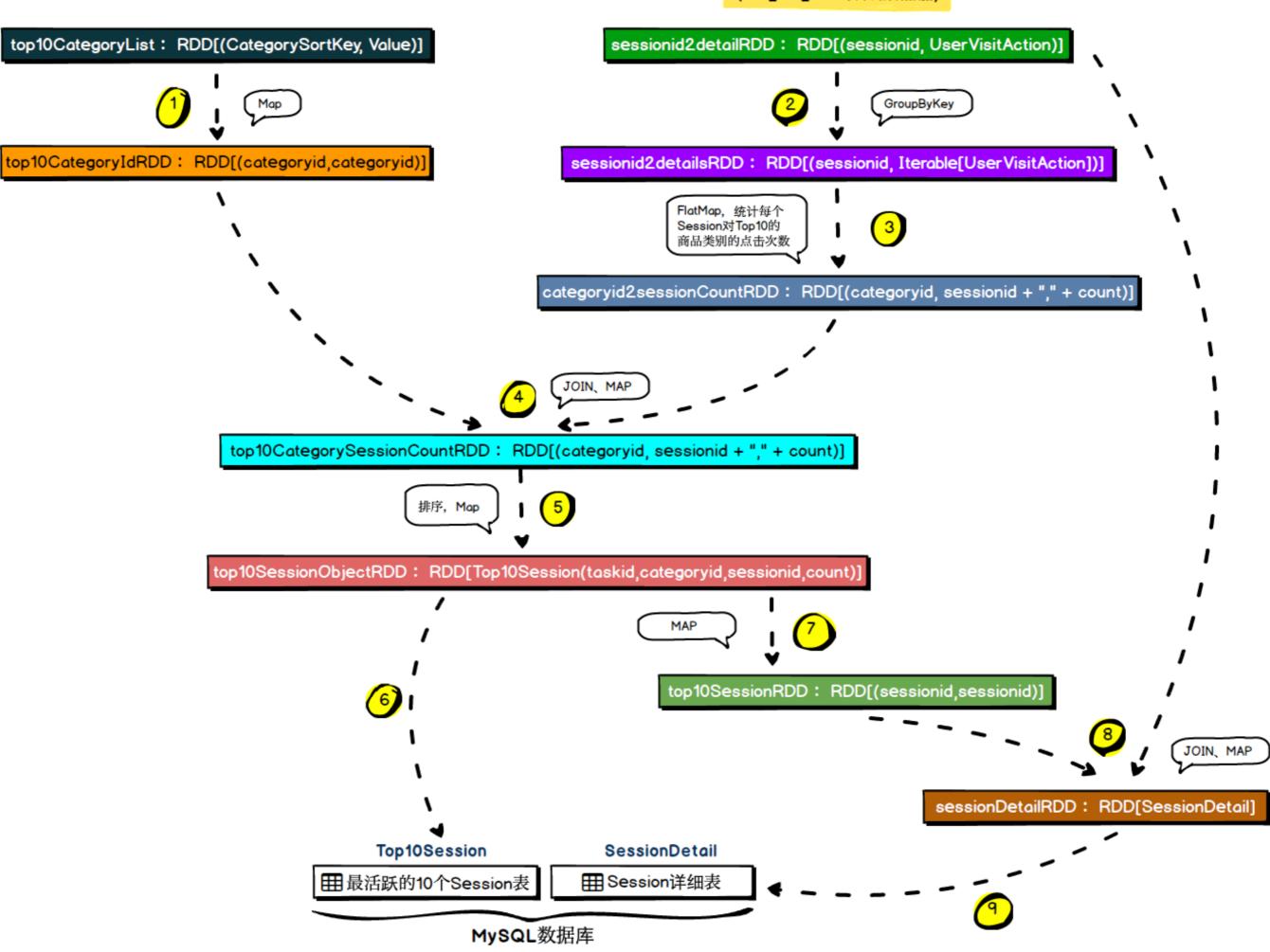
需求3: 获取Top10热门品类 7 / 14



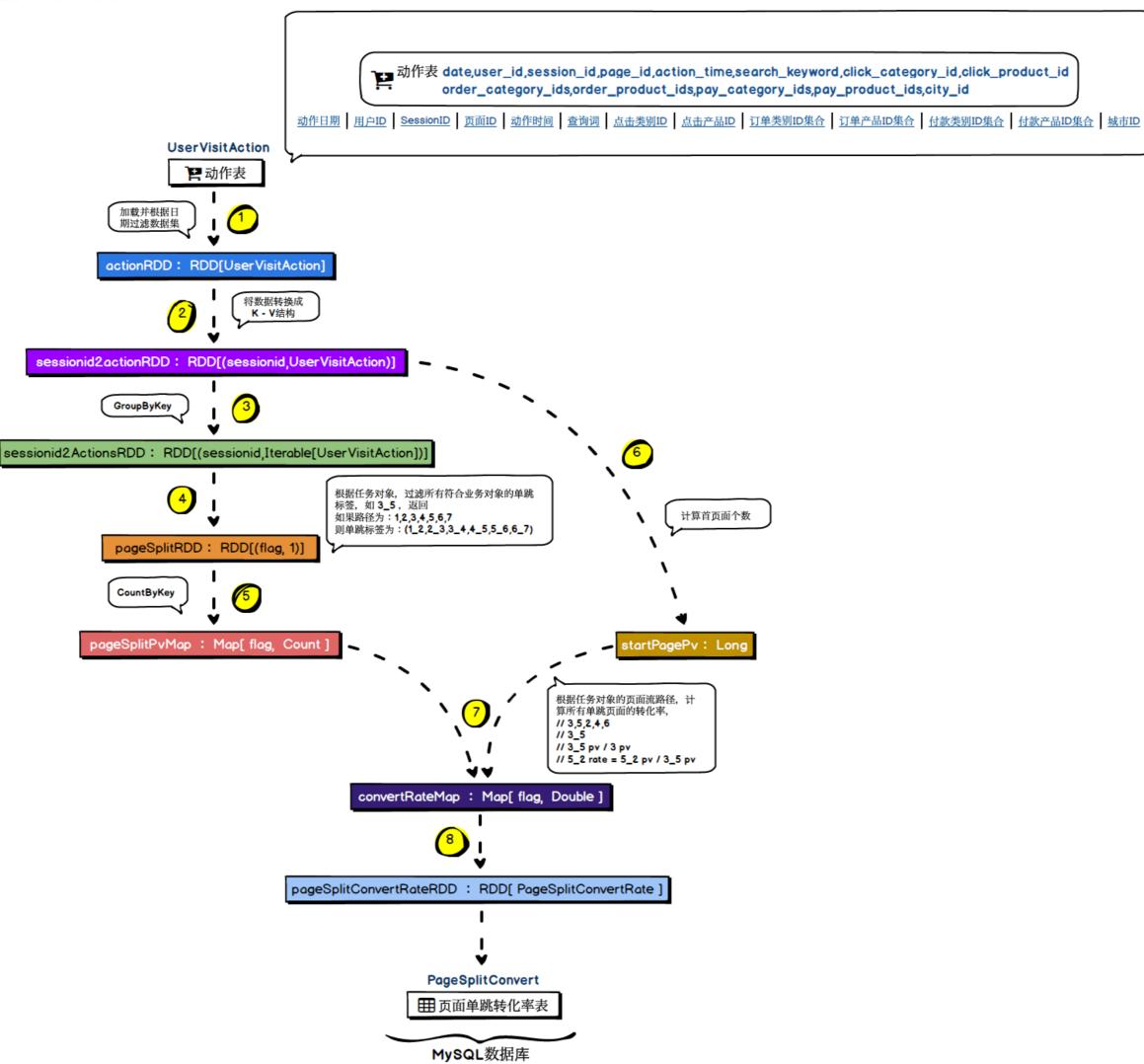


需求4: 获取Top10热门品类的活跃session

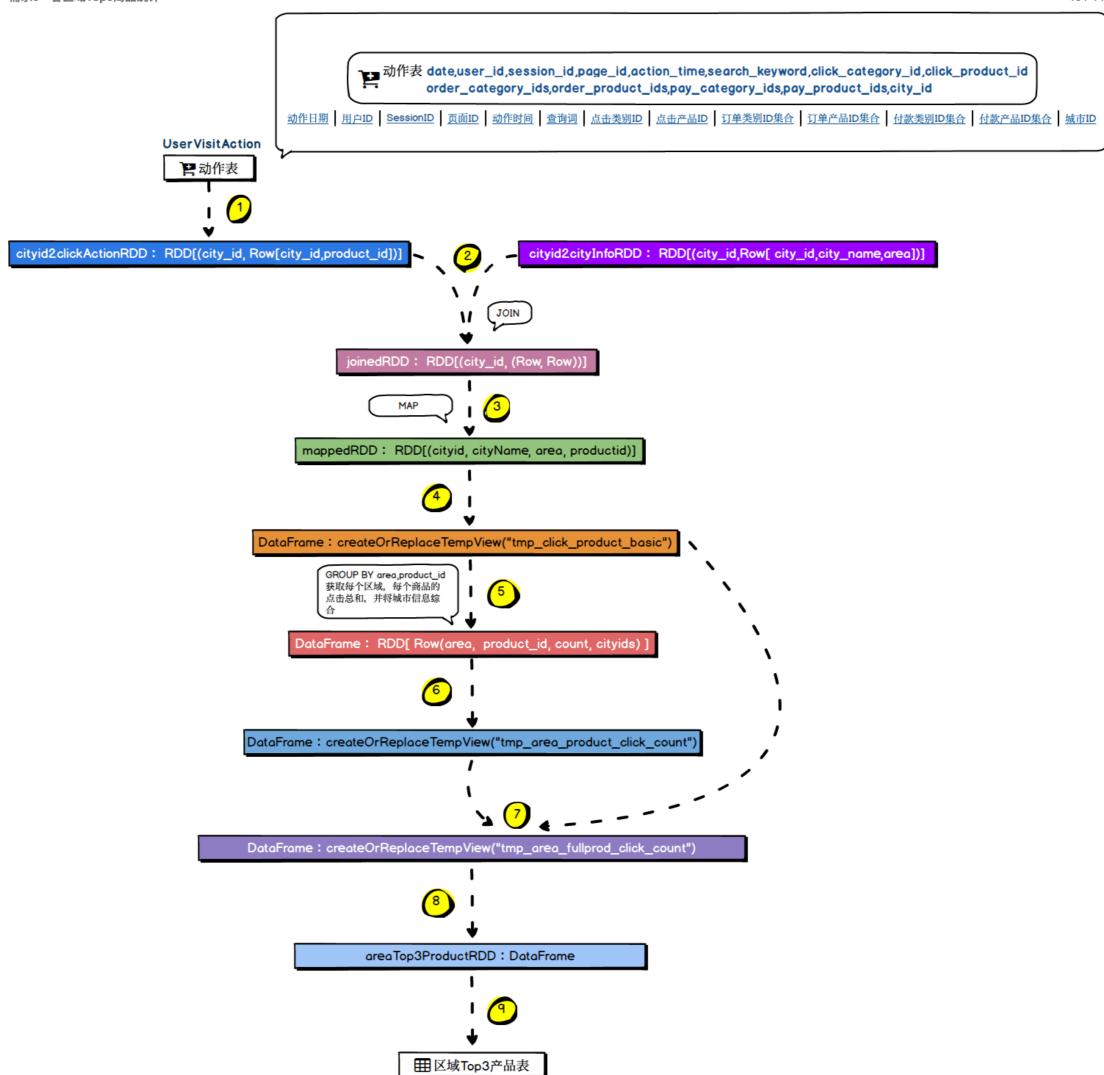


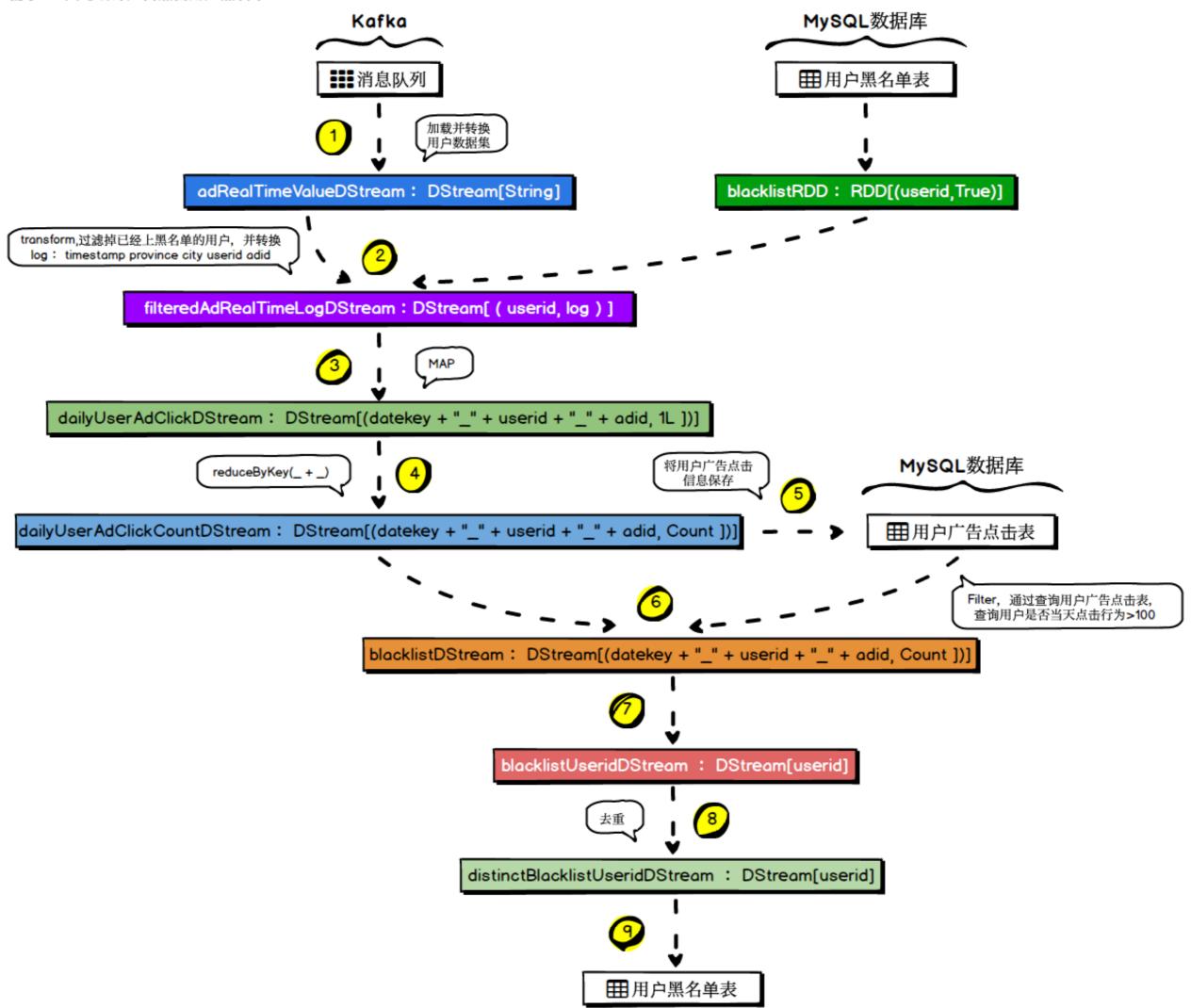


需求5: 页面单跳转化率统计 9 / 14



需求6: 各区域Top3商品统计 10 / 14





需求8: 实时统计广告点击量 12 / 14

