项目说明文档

一、需求说明

项目名称：xxx用户大数据观象台

需求概述：

用户session行为离线分析模块：

1. 通过指定taskid，从数据库查询任务相关信息，包括starttime，taskParam等，要求通过数据库连接池的方式获取链接，连接池在整个程序的运行过程中只有一份，且链接数量固定（Jdbc+Mysql）
2. 在指定日期范围内，按照session粒度进行数据聚合。要求聚合后的pair RDD的元素是<k:String,v:String>, 其中k=sessionid v的格式如下：

sessionid=value|searchword=value|clickcaterory=value|age=value|professional=value|city=value|sex=value（Spark RDD + Sql）

1. 根据用户的查询条件，一个或者多个：年龄范围，职业（多选），城市（多选），搜索词（多选），点击品类（多选）进行数据过滤，注意：session时间范围是必选的。返回的结果RDD元素格式同上（Spark RDD + Sql）
2. 实现自定义累加器完成多个聚合统计业务的计算，统计业务包括访问时长：1~3秒，4~6秒，7~9秒，10~30秒，30~60秒的session访问量统计，访问步长：1~3个页面，4~6个页面等步长的访问统计 注意：业务较为复杂,需要使用多个广播变量时，就会使得程序变得非常复杂，不便于扩展维护（Spark Accumulator）
3. 对通过筛选条件的session，按照各个品类的点击、下单和支付次数，降序排列，获取前10个热门品类。优先级：点击，下单，支付。二次排序（Spark）

广告流量实时分析模块需求：

1. 实现实时的动态黑名单机制，将每天对某个广告点击超过100次的用户拉黑
2. 基于黑名单的非法广告点击流量过滤
3. 统计每天各省各城市各广告的点击流量实时统计
4. 统计每天各省的top3热门广告
5. 统计各个广告最近一个小时内的点击趋势：各个广告最近1小时内各分钟的点击量，基于2
6. 实时计算每天各省城市各广告的点击量，更新到mysql

实现思路：

1、实时计算各batch中的每天各用户对各广告的点击次数

2、使用**高性能**方式将每天各用户对各广告的点击次数写入MySQL中（更新）

3、使用filter过滤出每天对某个广告点击超过100次的黑名单用户，并写入MySQL中

4、使用transform操作，对每个batch RDD进行处理，都动态加载MySQL中的黑名单生成RDD，然后进行join后，过滤掉batch RDD中的黑名单用户的广告点击行为

5、使用updateStateByKey操作，实时计算每天各省各城市各广告的点击量，并时候更新到MySQL

6、使用transform结合Spark SQL，统计每天各省份top3热门广告：首先以每天各省各城市各广告的点击量数据作为基础，首先统计出每天各省份各广告的点击量；然后启动一个异步子线程，使用Spark SQL动态将数据RDD转换为DataFrame后，注册为临时表；最后使用Spark SQL开窗函数，统计出各省份top3热门的广告，并更新到MySQL中

具体实现步骤：

* 1. 计算出每5秒的batch中，每天每个用户每个广告的点击数，对原始实时日志聚合，处理成<yyyyMMdd\_userid\_adid,1L>的格式
  2. 针对处理后的日志，执行reduceByKey算子，得到每天每个用户对广告的点击数量，并将结果写入(更新)到mysql
  3. 遍历每个batch中的所有记录，对每条记录都要查询一下，这一天这个用户对这个广告的点击量是多少，生成黑名单用户，写入mysql表，生成动态黑名单用户

**广告点击流量实时统计模块业务说明：**

网站 / app，是不是通常会给一些第三方的客户，打一些广告；也是一些互联网公司的核心收入来源；广告在网站 / app某个广告位打出去，在用户来使用网站 / app的时候，广告会显示出来；此时，有些用户可能就会去点击那个广告。

广告被点击以后，实际上，我们就是要针对这种用户行为（广告点击行为），实时数据，进行实时的大数据计算和统计。

每次点击一个广告以后，通常来说，网站 / app中都会有埋点（前端的应用中，比如JavaScript Ajax；app中的socket网络请求，往后台发送一条日志数据）；日志数据而言，通常，如果要做实时统计的话，那么就会通过某些方式将数据写入到分布式消息队列中（Kafka）；日志写入到后台web服务器（nginx），nginx产生的实时的不断增加 / 更新的本地日志文件，就会被日志监控客户端（比如flume agent），写入到消息队列中（kafka），我们要负责编写实时计算程序，去从消息队列中（kafka）去实时地拉取数据，然后对数据进行实时的计算和统计。

这个模块的意义在于，让产品经理、高管可以实时地掌握到公司打的各种广告的投放效果。以便于后期持续地对公司的广告投放相关的战略和策略，进行调整和优化；以期望获得最好的广告收益。

基于用户session行为的商品推荐模块：

1. 基于用户session访问相似度做离线推荐。
2. 基于商品购买的协同过滤，在购买某个商品的同时，购买次数最多的其他商品做在线推荐

数据量：

每天平均1TB ~ 25TB，节点19到157节点不等。

二、项目描述

**项目技术架构描述：**

1、J2EE的平台（美观的前端页面），通过这个J2EE平台可以让使用者，提交各种各样的分析任务，其中就包括一个模块，就是用户访问session分析模块；可以指定各种各样的筛选条件，比如年龄范围、职业、城市等等。。

2、J2EE平台接收到了执行统计分析任务的请求之后，会调用底层的封装了spark-submit的shell脚本（Runtime、Process），shell脚本进而提交我们编写的Spark作业。

**3、Spark作业获取使用者指定的筛选参数，然后运行复杂的作业逻辑，进行该模块的统计和分析。**

**4、Spark作业统计和分析的结果，会写入MySQL中，指定的表**

5、最后，J2EE平台，使用者可以通过前端页面（美观），以表格、图表的形式展示和查看MySQL中存储的该统计分析任务的结果数据。

**用户访问session介绍：**

用户在电商网站上，通常会有很多的点击行为，首页通常都是进入首页；然后可能点击首页上的一些商品；点击首页上的一些品类；也可能随时在搜索框里面搜索关键词；还可能将一些商品加入购物车；对购物车中的多个商品下订单；最后对订单中的多个商品进行支付。用户的每一次操作，其实可以理解为一个action，比如点击、搜索、下单、支付用户session，指的就是，从用户第一次进入首页，session就开始了。然后在一定时间范围内，直到最后操作完（可能做了几十次、甚至上百次操作）。离开网站，关闭浏览器，或者长时间没有做操作；那么session就结束了。

以上用户在网站内的访问过程，就称之为一次session。简单理解，session就是某一天某一个时间段内，某个用户对网站从打开/进入，到做了大量操作，到最后关闭浏览器。的过程。就叫做session。

session实际上就是一个电商网站中最基本的数据和大数据。那么大数据，面向C端，也就是customer，消费者，用户端的，分析，基本是最基本的就是面向用户访问行为/用户访问session。

三、开发组

**项目组：**

1. 需求组（2人）

完成项目需求文档整理和模型文档整理。

2．Web组（3人）

完成以下两个功能模块的页面开发。

详单页面展现：通过thrift与Spark对接，完成通过时间段和指令查 询原始详单。

分析结果数据展现：通过JDBC查询任务，保存分析结果。

3、DBA（2人）

完成导入数据库的存储过程，复杂关联查询和数据库维护工作

4、大数据组（9人，2个核心成员，6个开发，1个顾问）

大数据处理，任务入库、查询，hive入库，sparksql实时查询日志上报等核心模块开发 。Oozie各个处理流程wf开发、调试、优化等。监控系统功能测试、性能测试等。GBase etl模块spark开发。集群搭建、参数调优、后期维护等工作。负责出具相关的部署、使用、维护文档，对现场一线人员进行培训。

5.C++组（2人）

负责采集程序开发调试。