## 对用户所访问的session数据按条件进行分析

根据使用者（需求方）指定的一些条件，筛选出指定的一些用户（比如按年龄，职业或所在城市等筛选条件）

按条件筛选session，这个功能最大的作用就是灵活。可以让使用者（决策者），对感兴趣的和关心的用户群体，进行后续各种复杂业务逻辑的统计和分析，比如查找搜索过某些关键词的用户、访问时间在某个时间段内的用户、年龄在某个范围内的用户、职业在某个范围内的用户、所在某个城市的用户，等等条件发起的session。拿到的结果数据，就是只是针对特殊用户群体的分析结果，而不是对所有用户进行分析的泛泛的分析结果。例如现在某个企业高层，他想看到用户群体中，28~35岁的消费群体对应的一些统计和分析的结果数据，从而辅助高管进行公司战略上的决策制定。

## 用户访问session进行聚合统计

比如需要统计访问时长在0~3的session占比，4~6的session占比。

session访问时长，就是一个session对应的开始的action到结束的action之间的时间范围。访问步长，指的是，一个session执行期间内，依次点击过多少个页面，比如说一次session维持了1分钟，那么访问时长就是1分钟，然后在这1分钟内点击了10个页面，那么session的访问步长就是10。

session占比：举个栗子，符合第1步筛选出来的session的数量大概是有1000万个。在这里面，我们要计算出访问时长在1s~3s内的session的数量，并除以符合条件的总session数量（比如1000万），比如是100万/1000万，那么1s~3s内的session占比就是10%，依次类推。

这个功能的作用，其实就是，可以让人从全局的角度看到，符合某些条件的用户群体使用我们的产品的一些习惯。比如大多数人，到底是会在产品中停留多长时间，大多数人会在一次使用产品的过程中访问多少个页面。那么对于决策者来说，有一个清晰的全局认识。

## 按时间比例随机抽取session

随机抽取本身是很简单的，但是按照时间比例，就复杂了。比如说，这一天总共有1000万的session。现在要从这1000万session中随机抽取出来1000个session。但是这个随机不是那么简单的。需要做到如下要求：首先，如果这一天的12:00~13:00的session数量是100万，那么这个小时的session占比就是1/10，那么这个小时中的100万的session，我们就要抽取1/10 \* 1000 = 100个。然后再从这个小时的100万session中，随机抽取出100个session。以此类推，其他小时的抽取也是这样做。

这个功能的作用是可以让使用者能够对于符合条件的session按照时间比例均匀的随机采样出1000个session，然后观察每个session具体的点击行为，比如先进入了首页，然后点击了食品品类，接下来点击了雨润火腿肠商品，然后搜索了火腿肠罐头的关键词，接着对王中王火腿肠下了订单，最后对订单做了支付。

之所以要做到按时间比例随机采用抽取，就是要做到，观察样本的公平性。

## 获取点击、下单、支付次数排名的top10

计算出来通过筛选条件的session，他们访问的所有品类（点击、下单、支付），按照各个品类的点击、下单、支付次数，进行降序排列，然后再获取前10个品类，也就是筛选条件下的那一批session的top10热门品类。

怎么排序？点击、下单、支付次数：优先按照点击次数排序，如果点击次数相等，就按照下单次数来排序，以此类推，这就用到了自定义排序。

对session按照条件排序，用到二次排序（自定义排序）

这个功能很重要，可以让我们了解符合条件的用户最感兴趣的商品是什么品类。这个可以让公司里的相关负责人，清晰地了解到不同层次、不同类型的用户的购买习惯。

## top10活跃session

从top10热门品类，获取每个品类点击次数最多的前10个session，以及对应的访问明细。

这个功能可以分析出每个品类中的最受欢迎或最受关注的商品。

## 页面单跳转化率

1. 接收J2EE系统传进来的taskId，根据taskId去MySQL查询任务的参数，日期范围、页面流id
2. 针对指定范围日期内的用户访问行为数据，去判断计算页面流id中每两个页面组成的页面切片，它的访问量是多少
3. 根据指定页面流中各个页面切片的访问量，计算出各个页面切片的转化率
4. 把计算出来的转化率，写入MySQL

## 区域热门top3商品

根据用户指定的日期范围，来统计各个区域下的最热门的top3商品

Spark作业接收taskId，查询对应的MySQL中的task，获取用户指定的筛选参数，进而统计出指定日期范围内的各个区域top3热门商品，最后将结果写入数据库

## 广告点击流量实时统计

1. 实时的动态获取黑名单机制，将每天对每个广告点击超过100次的用户拉黑
2. 基于黑名单的非法广告点击流量过滤机制
3. 每天各省各城市各广告的点击流量实时统计，基于第二点的数据基础之上
4. 统计每天各省top3热门广告
5. 统计各广告最近1小时内的点击量趋势，各广告最近1小时内各分钟的点击量，也是基于第2点的数据基础之上
6. 使用高性能方式将实时统计结果写入MySQL

## 9、数据分析优化方案（性能调优）