京东个性化推荐技术实践

推荐搜索部　王志勇

2014-09-27

JD.京东.COM

目录

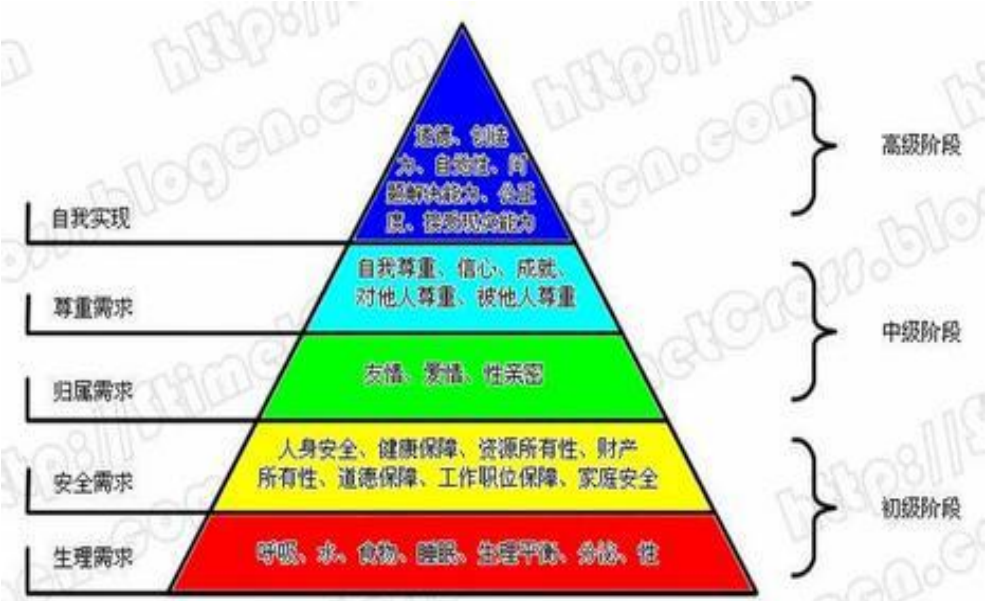
- 个性化推荐简介
  - 为什么需要个性化推荐
  - 个性化推荐为什么需要实时
- 个性化推荐系统的实践
  - 个性化推荐面临的问题
  - 京东个性化推荐的实践方案
  - 应用与效果分析

Personalized Recommendations

个性化推荐简介

用户需求的多样性

– 幸福，就是自己的需求被满足



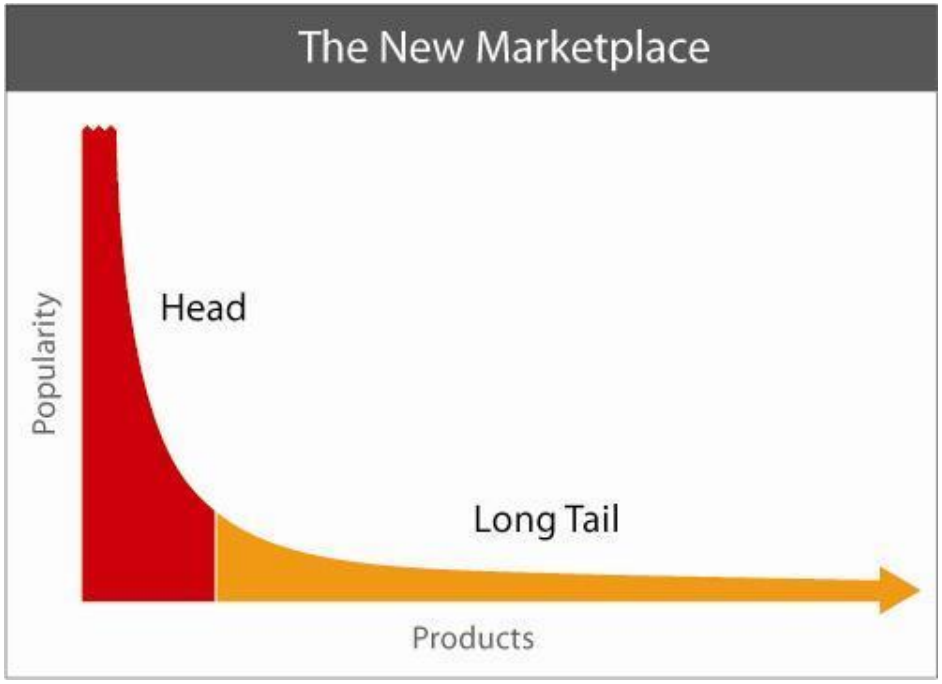
马斯洛需求层次理论（Maslow's hierarchy of needs）

信息过载 VS 用户注意力资源

京东有近200,000,000商品可选购， 每天新增商品超过500,000

长尾理论 --- 个性化时代的到来

- 突破丰富的有限
- 突破传统的生产成本和流通成本



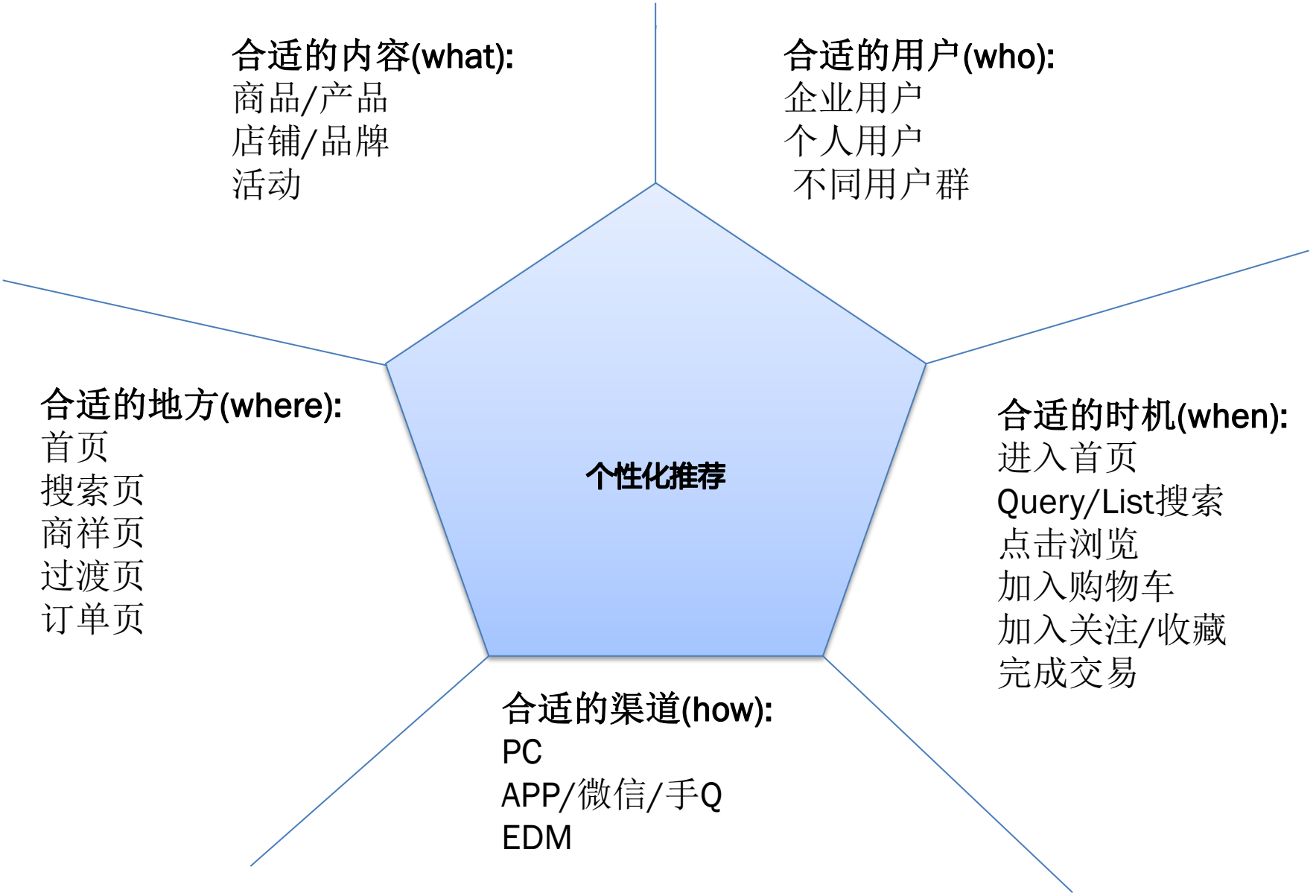
个性化推荐的作用

提高用户忠诚度和用户体验，提高用户购物决策的质量和效率

提高成交转化率(CTR,CVR,GMV)

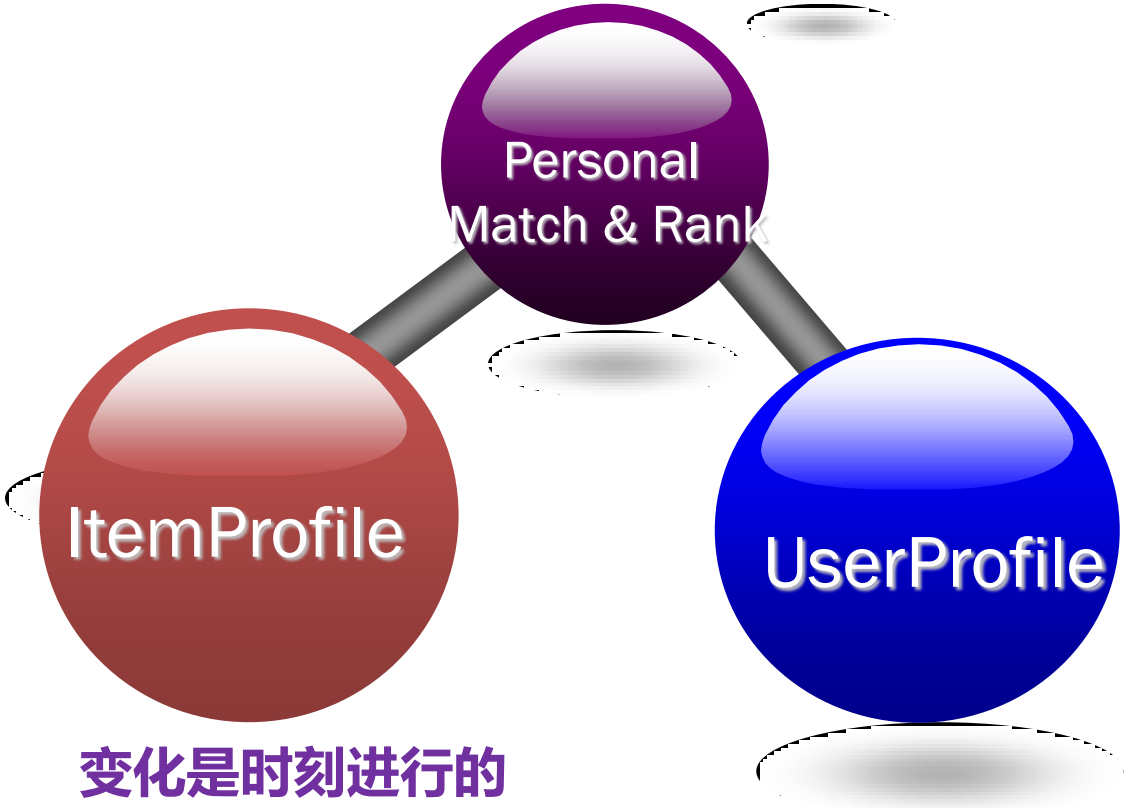
提高网站交叉销售能力

个性化推荐需要解决的问题



个性化为什么需要实时

- 个性化因素的变化



Item changing overtime

User changing overtime

变化是时刻进行的

商品在变化

用户个体在变化

群体、环境在变化

个体和群体的隶属关系也在动态变化

Personalized Recommendations

个性化推荐实践

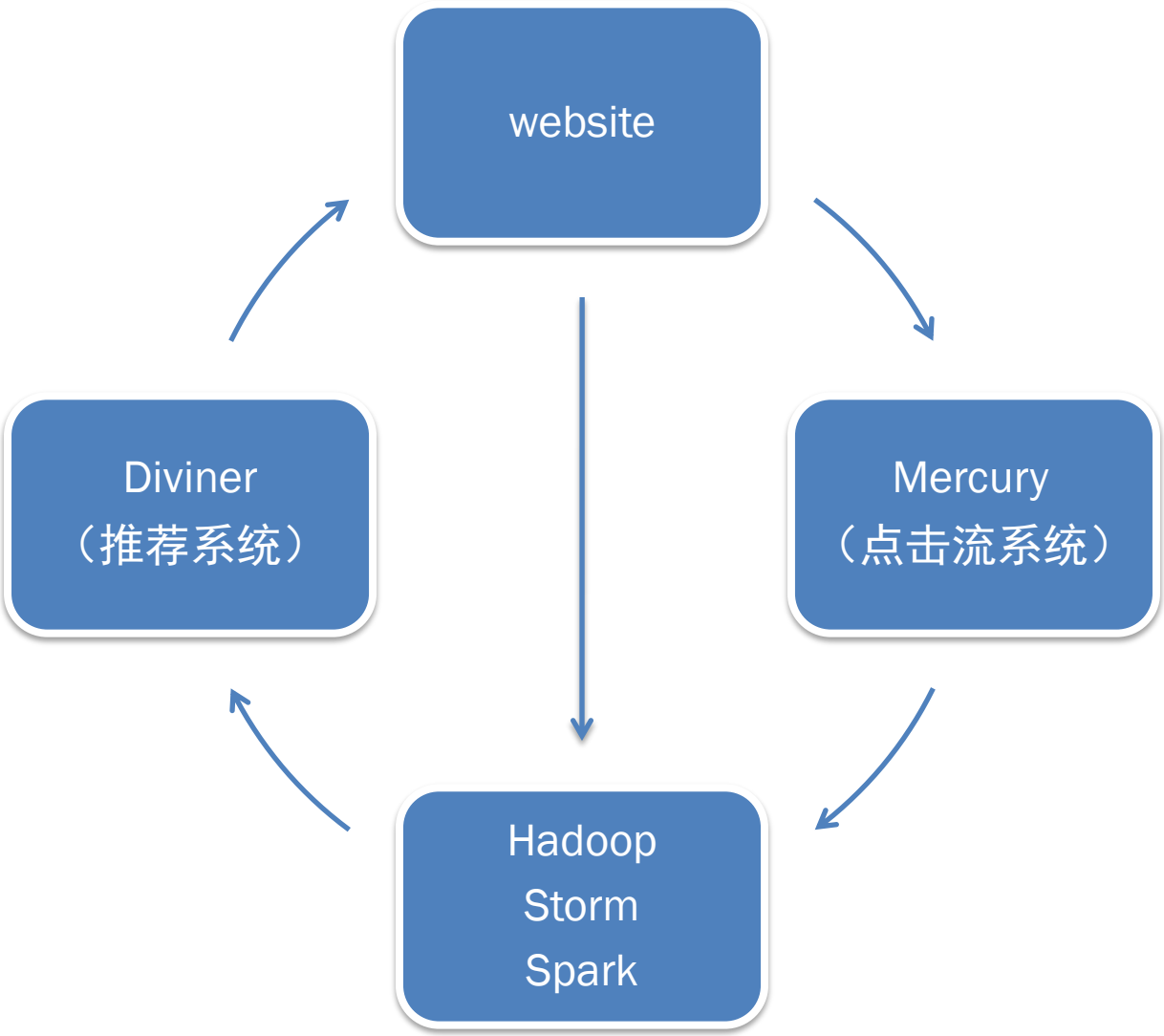
个性化推荐三大挑战

- 处理持续增长的大数据能力
- 实时分析用户购物意图能力
- 大规模稀疏数据建模能力

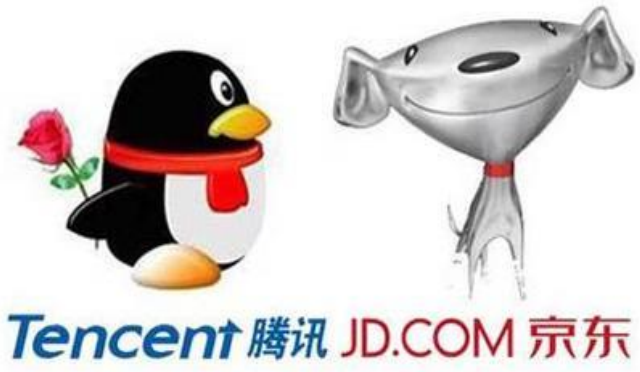
京东实践之路

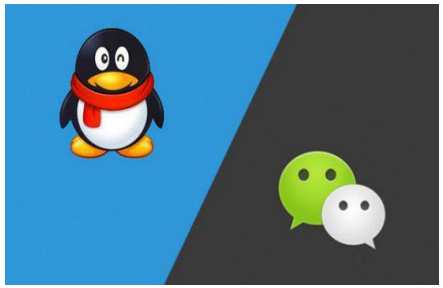
- 数据收集/存储
- 离线/在线分析
  - 数据清洗
  - 数据建模
    - 商品建模
    - 用户建模
- 在线推荐
- A/B测试平台

数据收集(一)　大数据存储，实时计算



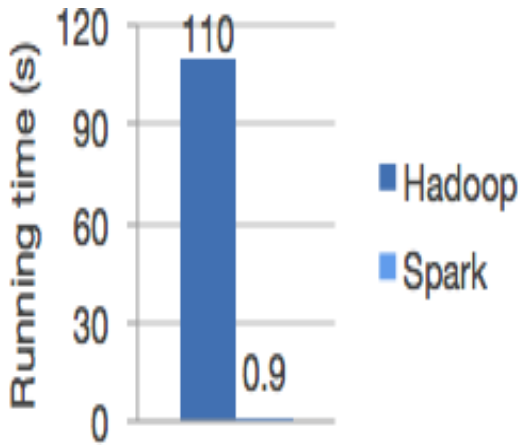
数据收集(二)　　用户数据打通

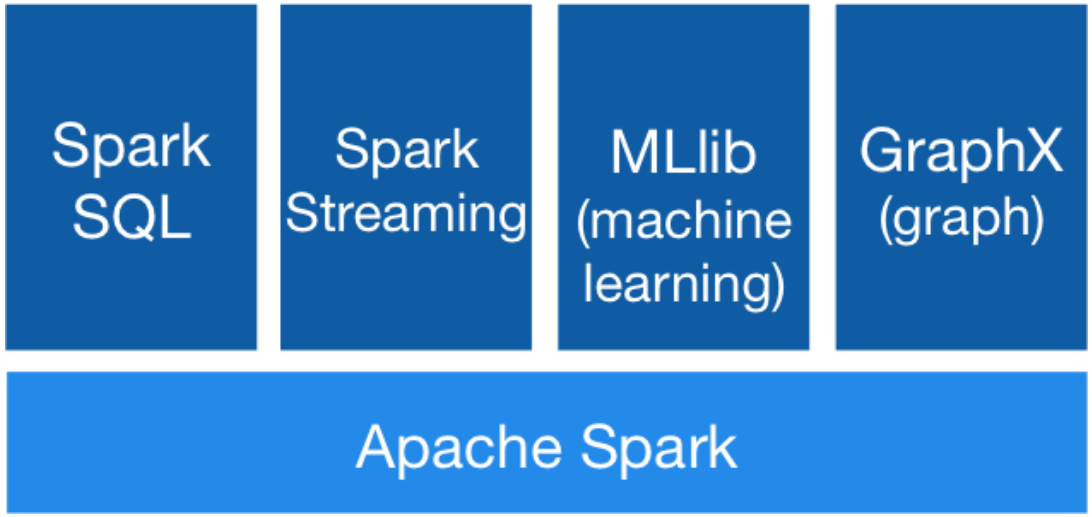




离线计算框架

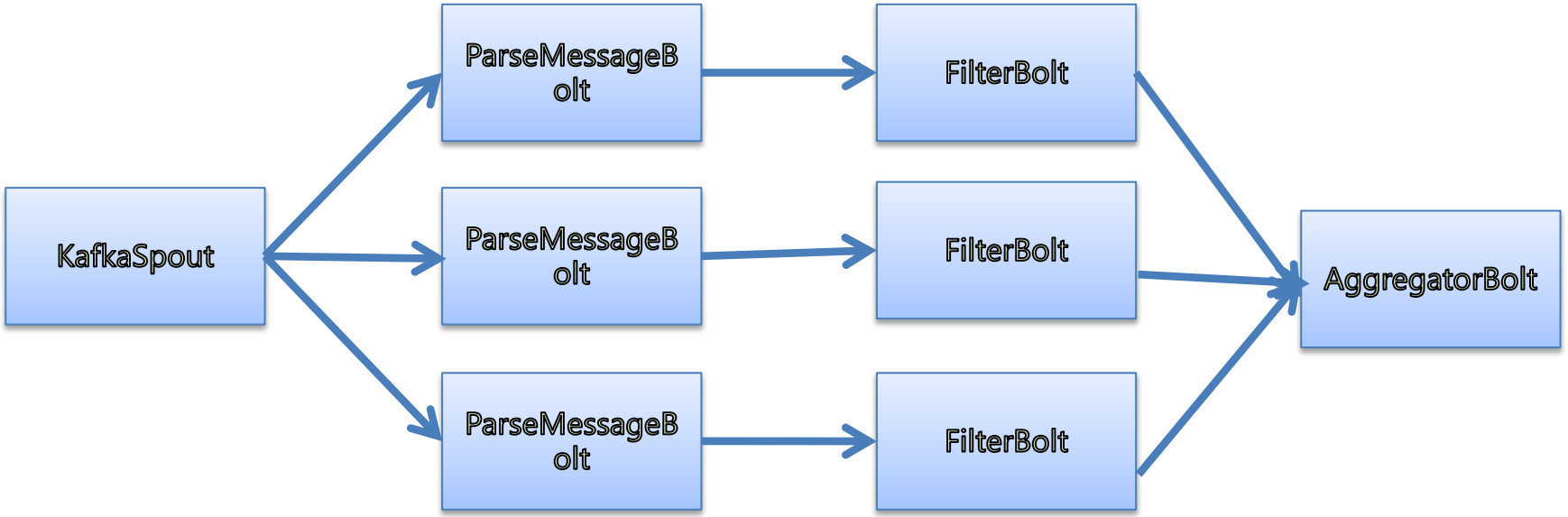
- 离线计算主要在**Hadoop**上运行**Map Reduce**
- 部分计算都使用**Mahout，Spark**





在线计算框架

- 在线计算主要基于Storm，实时消息基于kafka(30亿+)。



数据清洗

- 数据为王，垃圾进垃圾出
- 现实数据是“肮脏”的
  - 数据残缺
  - 数据重复
  - 数据不一致
- 商业数据更“肮脏”
  - 点击作弊
  - 订单作弊
  - 交易作弊
  - 评论作弊……

数据建模(一) ---- 商品画像

属性

- 品牌
- 产品
- 颜色/尺码
- 风格/材质
- 适用人群
- 图片。。。

Apple苹果 全掌气垫运动鞋 男跑步鞋情侣鞋透气休闲鞋 土豪系列AP-5S 情侣款 黑绿桔红

流量

- PV
- UV
- CTR
- Load Time
- Exit Rate。。。

销售

- 销量/销售额
- 价格指数
- 促销类型
- 评论
- 退/换货率
- 物流速度
- 售后服务

商品画像挖掘算法

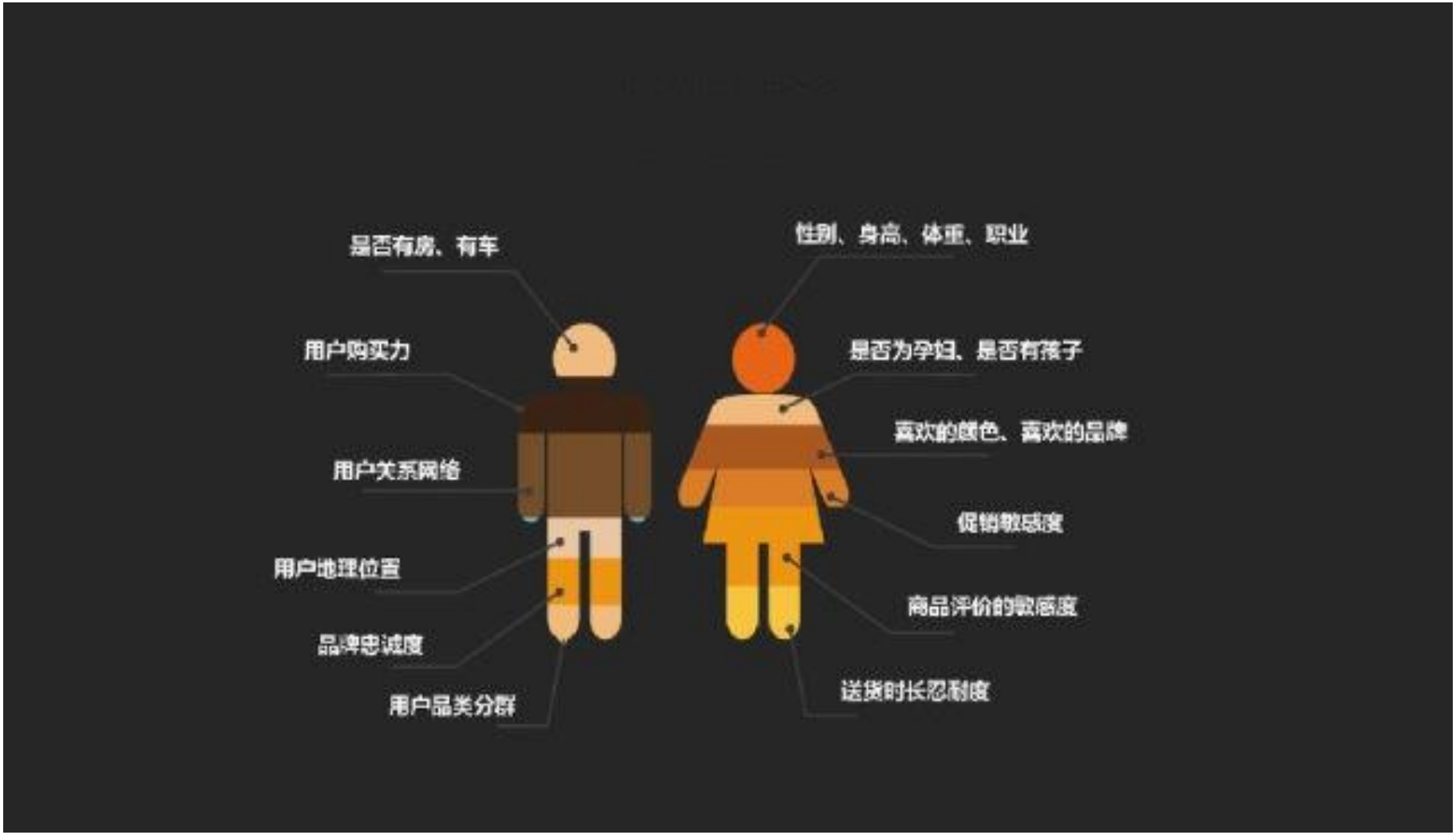
- **相似商品挖掘：**
  - **基于内容**
    - **LDA**
    - **SimHash**
  - **基于用户行为**
    - **Session浏览商品的CF**
    - **……**
- **相关商品挖掘：**
  - **基于商品的FP-Growth**
  - **基于产品的FP-Growth**
  - **基于图扩展**
  - **……**

商品画像服务(Clerk)

- **四大挑战**
  - 商品量大　--- **2.0亿 +**
  - 并发量大　--- **QPS 1.5~2万**， 日请求约**10亿**
  - 易变数据实时性 --- 库存，价格变化频繁
  - 性能要求高 ---- **TP999 < 10ms**

- **解决方案 ---　京东缓存云**
- ShardedJedis+ShardedJedisPool
  - 自动管理连接池，支持多个分片的独立连接池；
  - MasterFirst + MasterOnly

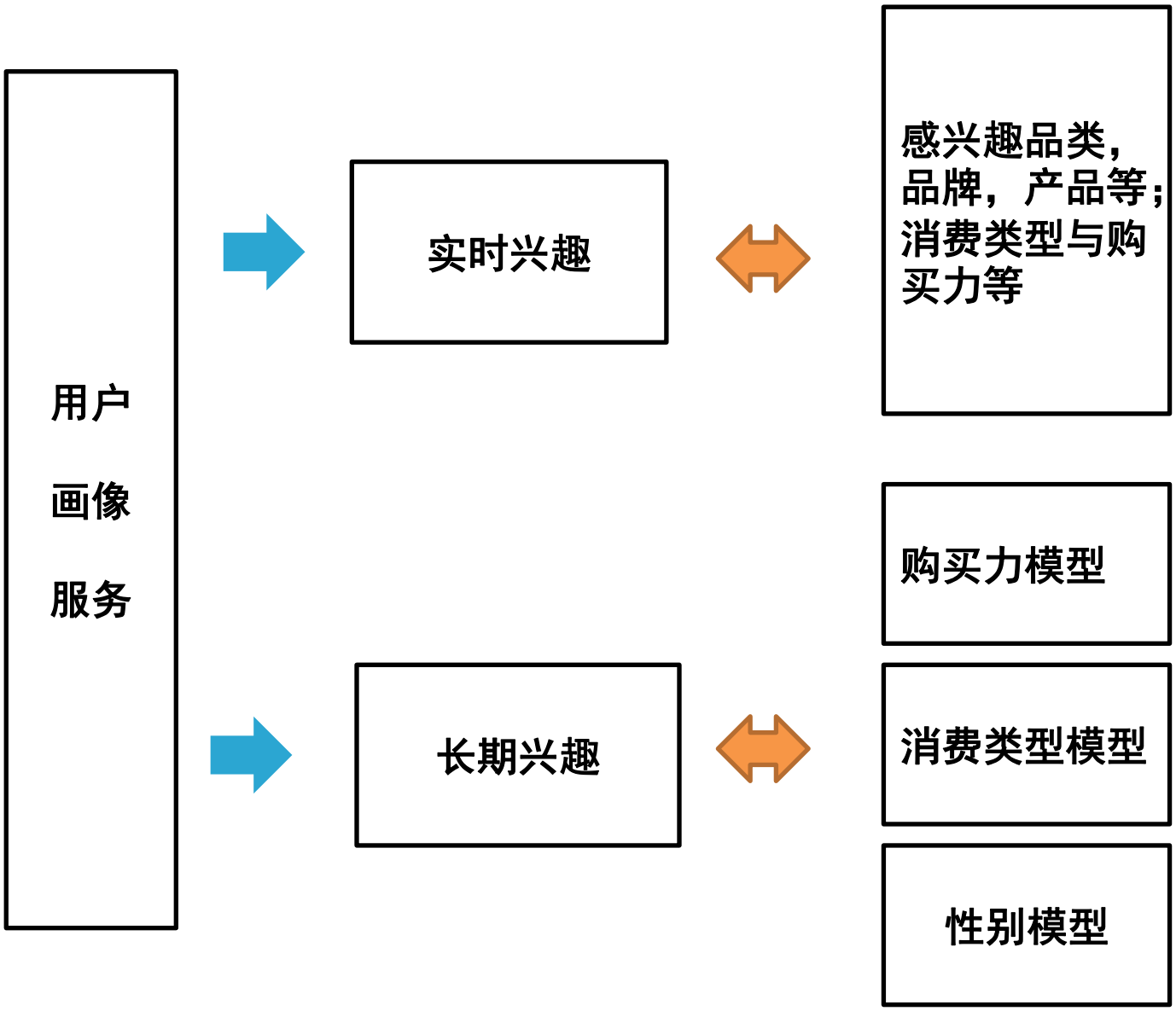
数据建模(二) ---- 用户画像


是否有房、有车
用户购买力
用户关系网络
用户地理位置
品牌忠诚度
用户品类分群
性别、身高、体重、职业
是否为孕妇、是否有孩子
喜欢的颜色、喜欢的品牌
促销敏感度
商品评价的敏感度
送货时长忍耐度

用户画像挖掘算法

- **用户品牌兴趣度模型：**
  - **GBDT + RankSVM**
- **用户购买力模型：**
  - **基于Mahout的ALS-WR**
- **用户品类兴趣度模型：**
  - **基于规则 优于 SVD++**
- **……**

用户画像服务(Zeus)



用户实时兴趣

Model Test

Tasks

用户pin		
fyw004		
三级分类	**三级分类**	**权值**
1396	套装	0.083
6286	立体拼插	0.083
6280	积木	0.083
品牌	**品牌**	**权值**
1396:12278	套装:曼秀雷敦（Mentholatum）	0.083
6286:11116	立体拼插:乐高（LEGO）	0.083
6280:11116	积木:乐高（LEGO）	0.083
产品关键词	**产品关键词**	**权值**
1396:49406	套装:洁面膏	0.083

用户长期兴趣

Display Server　Figure　Display

fyw004　查询　显示字段描述　显示服务器反馈信息

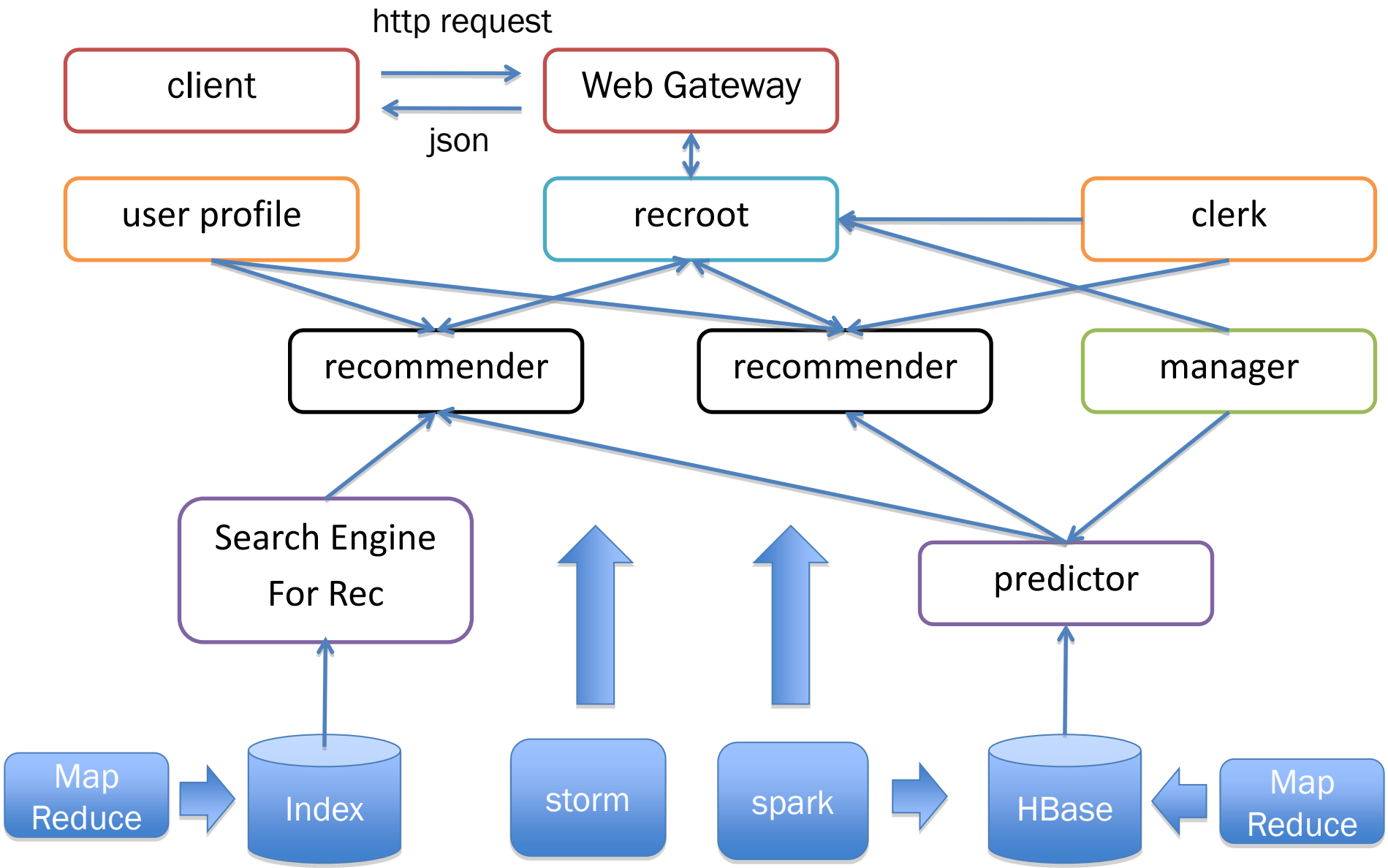
表名:portal_pin,列名:df.dm_user_model

```
userKey {
  serviceName: "portal_pin"
  id: "fyw004"
}
userModelProto {
  demography {
    用户名: "fyw004"
    是否有小孩: true
    孩子性别得分: BOY_GIRL
    性别: FEMALE
    性别置信度:high,medium,low: "high"
    生命周期类型: DECLINE
    用户价值分组: VG_VERY_HIGH
    标准化价值得分: 97
    用户下单最多的省份: "北京"
  }
  shoppingFeatures {
    用户购物类型: RATIONAL
    用户促销敏感度: HIGH
    关注商品评价模型: HIGH
    颜色偏好top1: WHITE
    颜色偏好top2: BLUE
    用户购买订单里最多的一级品类名称: "服饰鞋帽"
    用户最偏好的三个品牌: "U.S.POLO  ASSN.,O.SA,歌莉娅"
    用户品类分群模型: SUPERUSER
    用户忠诚度: LOYALTY
    单品促销敏感度: MIDDLE
    套装促销敏感度: LOW
    团购优惠促销敏感度: LOW
    满返满送促销敏感度: MIDDLE
    用户下单总数: 80
    局域网下单总数: 2
    网吧下单总数: 0
    学校下单总数: 0
    单位下单总数: 53
    家里下单总数: 25
    用户活跃度模型: VERY_ACTIVE
    大家电模型: true
    用户历史购买过订单里: 223
    购买力分段: 1-5从高到低: RICH
  }
}
```

表名:portal_pin,列名:df.search_cid3_interest

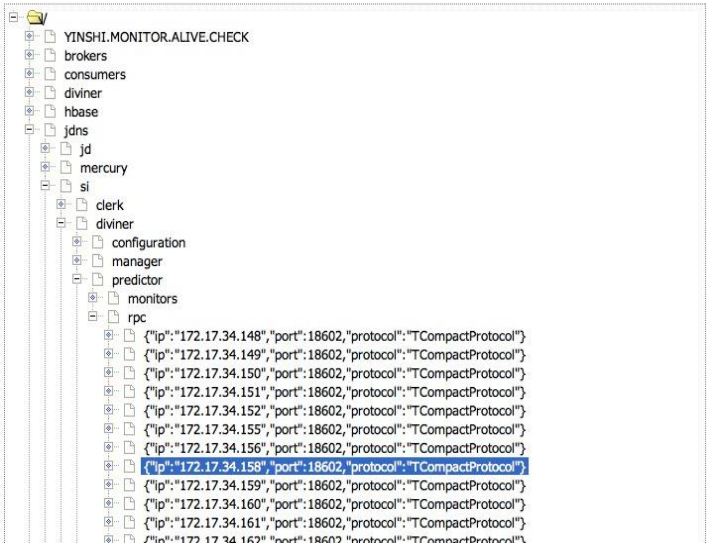
```
userKey {
  serviceName: "portal_pin"
  id: "fyw004"
}
userCid3ProperProto {
  用户名: "fyw004"
  用户偏好的三级分类 {
    proper: "11224"裤子
    weight: 100
  }
  用户偏好的三级分类 {
    proper: "11232"凉鞋
    weight: 65
  }
  用户偏好的三级分类 {
    proper: "9249"电炖锅
    weight: 52
  }
  用户偏好的三级分类 {
    proper: "1391"护肤
    weight: 37
  }
  用户偏好的三级分类 {
    proper: "11222"套装
    weight: 34
  }
  用户偏好的三级分类 {
    proper: "9775"拖鞋/人字拖
    weight: 33
  }
  用户偏好的三级分类 {
    proper: "1662"衣物清洁
    weight: 28
  }
  用户偏好的三级分类 {
    proper: "753"电饭煲
    weight: 22
  }
  用户三级分类下偏好的品牌词 {
    proper: "11224:19048"裤子:博士蛙
    weight: 100
  }
```

在线推荐系统架构


http request
client
Web Gateway
json
user profile
recroot
clerk
recommender
recommender
manager
Search Engine
For Rec
predictor
Map
Reduce
Index
storm
spark
HBase
Map
Reduce

服务基础设施及服务治理

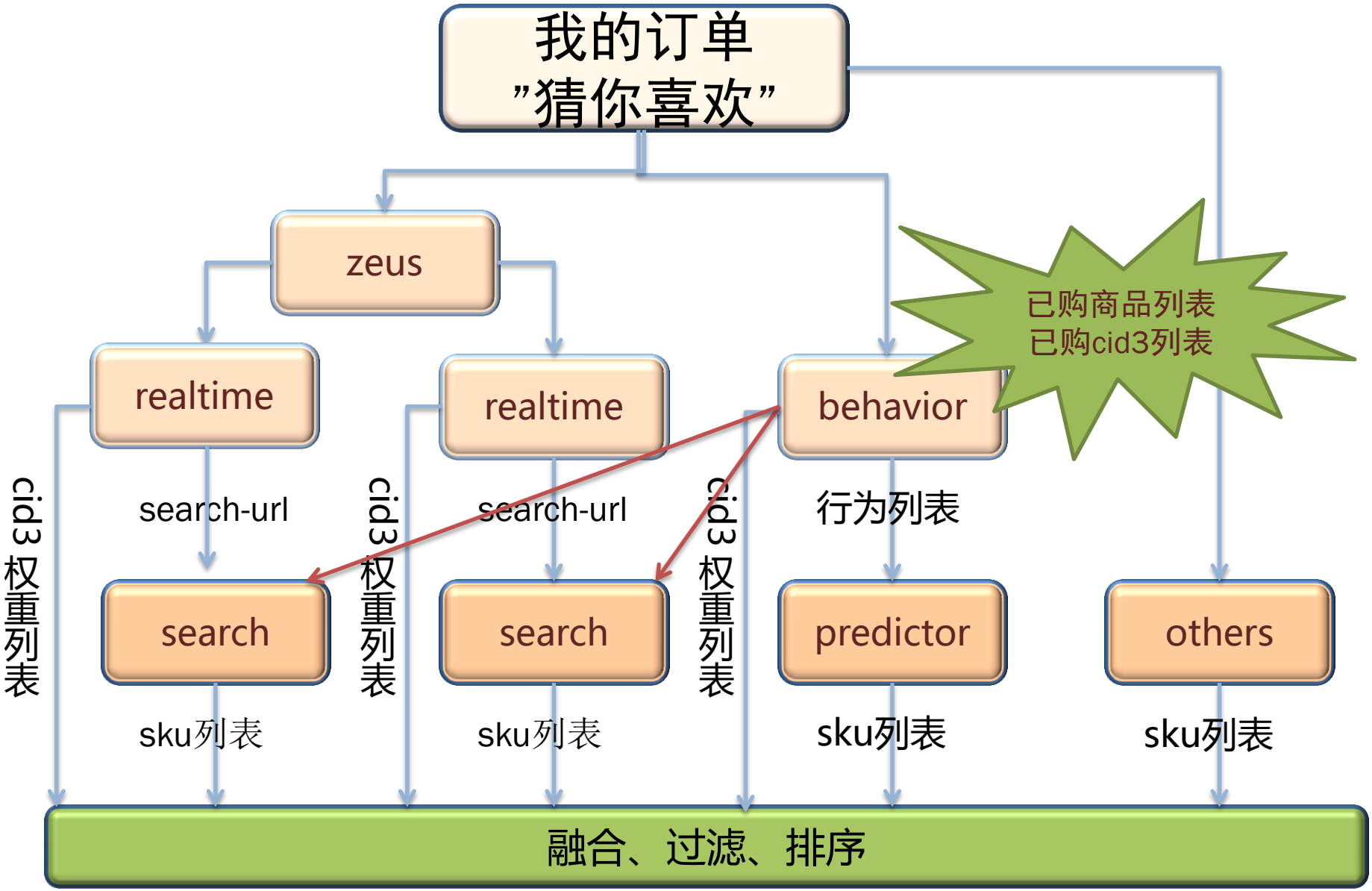
- use **Thrift** as RPC framework
  - multi-language support: Java / C++ / PHP
- use **zookeeper** to register replicated services
  - never interrupt
- **JDNS** (JD naming service)
  - service discovery
- multi-protocol support
- SAF compatible
- high performance / robust



在线算法主体流程

- 召回(recall)
  - 从一个或者多个Predictor中读取候选推荐目标
  - 从Search中召回候选推荐目标
- 过滤(filter)
  - 去掉无货、下架、用户已购商品
- 计分(rank)
  - 多模型融合
  - 在线CTR预估
- 排序(sort)
  - 对N个结果进行排序，取最前面的n个
- 填充(fill)
  - 查询clerk服务，将商品标题、价格、图片等信息补全

个性化推荐案例


我的订单
”猜你喜欢”
zeus
realtime
realtime
behavior
已购商品列表
已购cid3列表
search-url
search-url
行为列表
cid3 权重列表
cid3 权重列表
cid3 权重列表
search
search
predictor
others
sku列表
sku列表
sku列表
sku列表
融合、过滤、排序

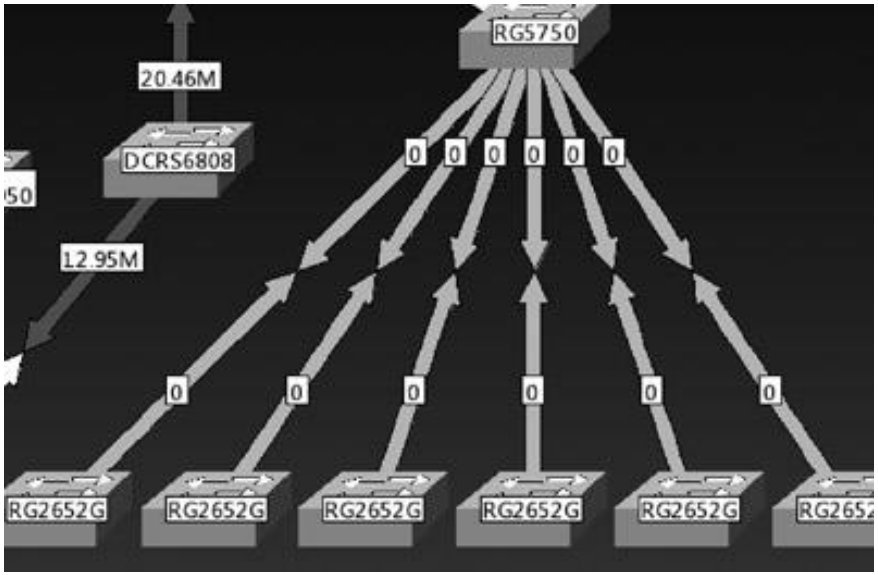
京东推荐A/B测试平台

A/B Test

- 无数据，不优化
  - 线上分流实验是进行推荐算法优化的必由之路
- 指标定义
  - CTR
  - 转化率
  - 销售额 (RPM ...)
- 数据收集
  - 点击流系统 (Mercury)
  - 数字签名，防篡改
- 数据处理：
  - Hive (standing query)
  - Shark (ad hoc query)
- 实验管理平台
- 实验报告

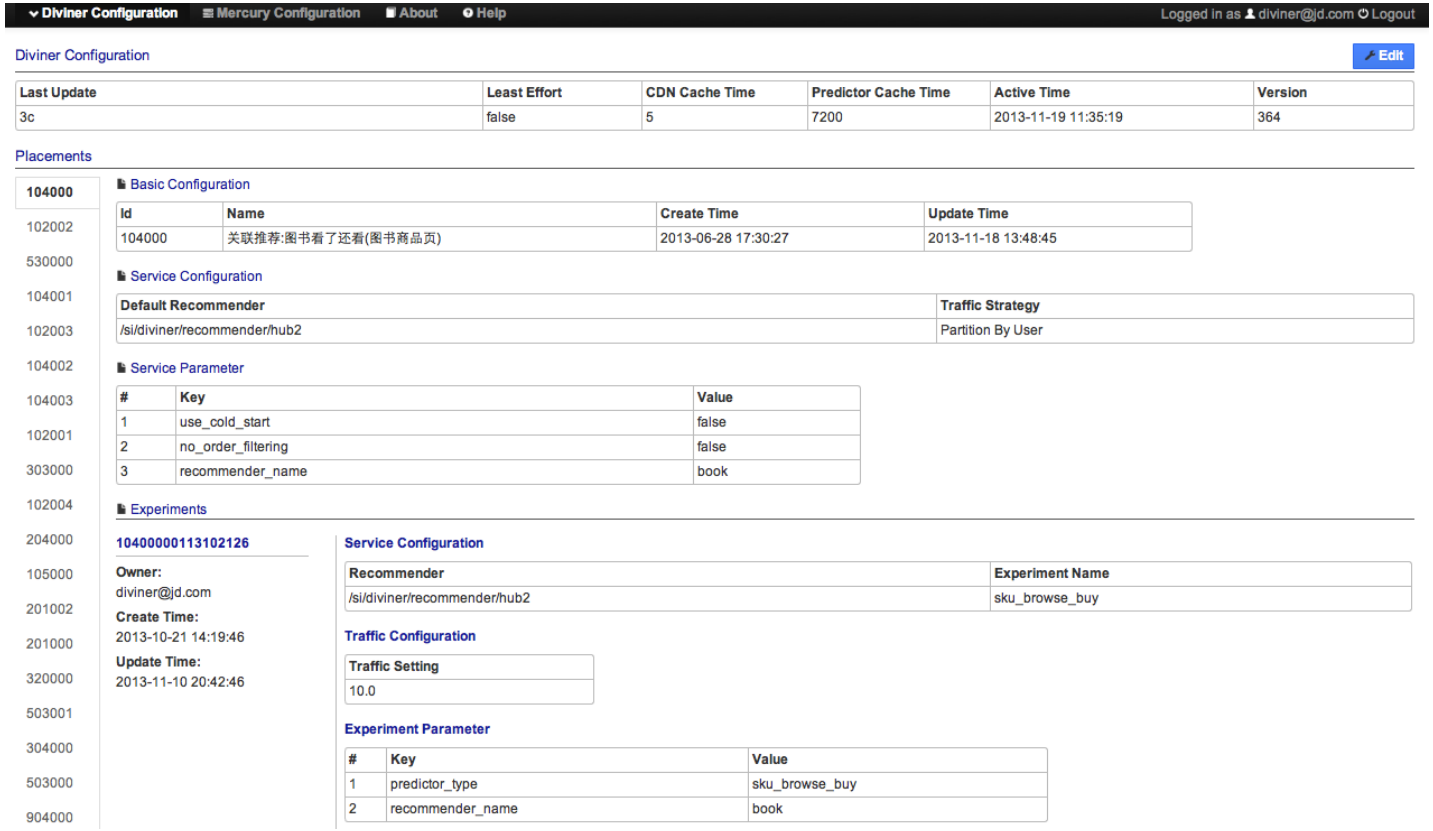
分流策略

- Random
  - 随机分流，用于可变结果集
- Partition By User
  - 按用户切分，同一用户永远看到同样结果
- Partition By Category
  - 按分类切分，针对不同分类测试算法针对性



配置管理后台

- 图形界面配置线上分流实验，配置即时生效，避免反复上线
- 增加实验，选择分流策略
- 调整流量，设置生效时间
- 配置追溯和回滚

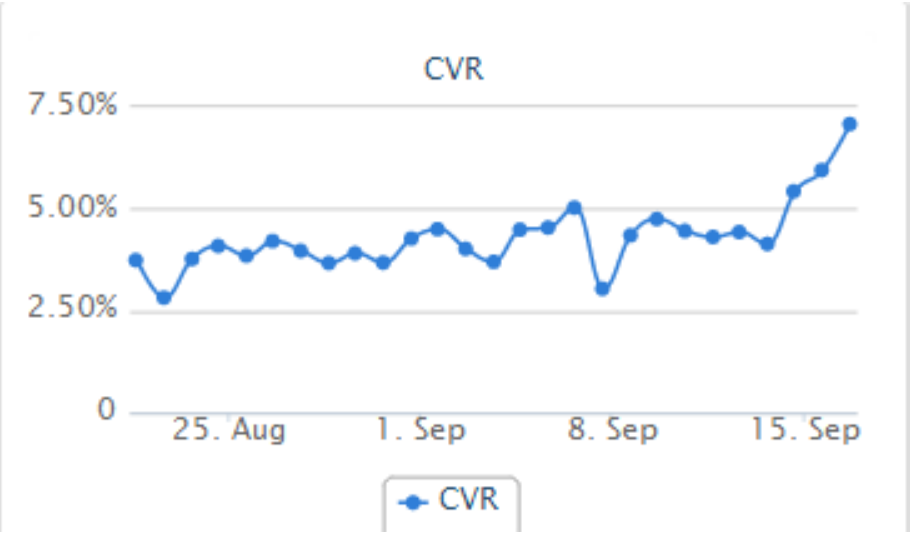


推荐结果持续改进

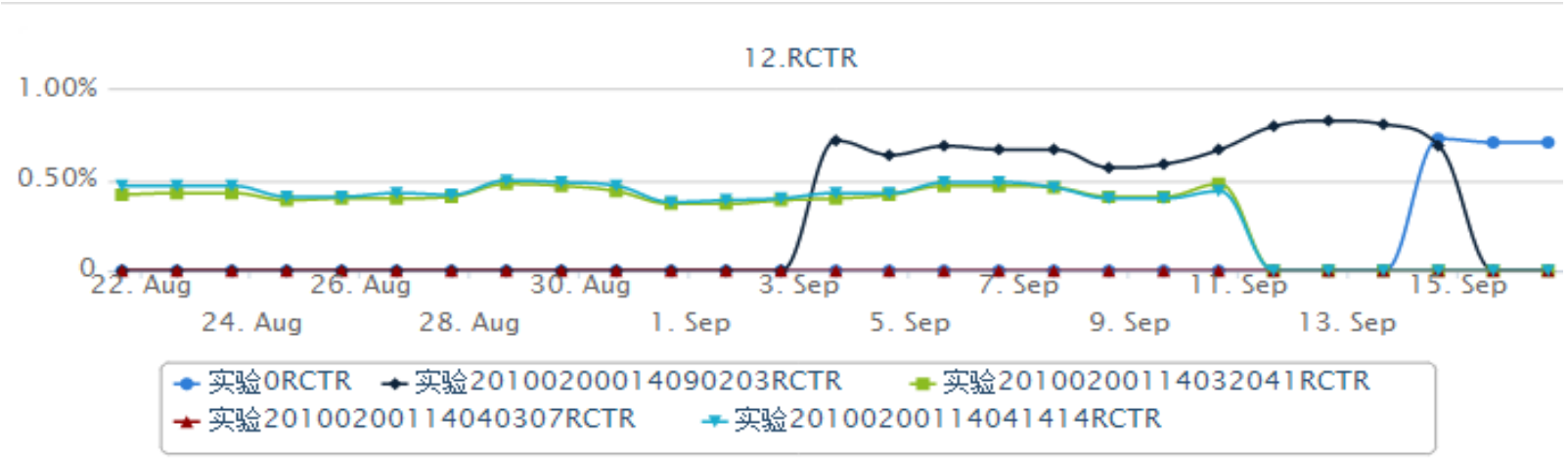
- 利用调试页面，对推荐结果进行内部review
- 将调试页面提供给产品或者业务review推荐结果，给出反馈意见
- 收到反馈意见后，与算法工程师讨论，安排算法优化
- 将优化后的结果进行线上分流实验，并将新结果调试页面发给产品或者业务，供其参照对比
- 收集实验数据，验证算法优化对指标的影响。将实验结果发送给产品或者业务，并解释算法优化的逻辑
- 优化算法正式上线
- 调试页面地址：

  http://diviner.jd.com/diviner?p=104001&uuid=1&sku=10490386&lid=1&lim=8&ec=gbk&fmt=dbg

个性化推荐案例效果


CVR
7.50%
5.00%
2.50%
0
25. Aug
1. Sep
8. Sep
15. Sep
CVR


订单行
3000
2000
1000
0
25. Aug
1. Sep
8. Sep
15. Sep
订单行


12.RCTR
1.00%
0.50%
0
22. Aug
24. Aug
26. Aug
28. Aug
30. Aug
1. Sep
3. Sep
5. Sep
7. Sep
9. Sep
11. Sep
13. Sep
15. Sep
实验0RCTR
实验201002000140902O3RCTR
实验20100200114032041RCTR
实验20100200114040307RCTR
实验20100200114041414RCTR

谢谢！