



全球软件案例研究峰会

100

TOP 100 CASE STUDIES OF THE YEAR

全球软件案例研究峰会



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com



全球软件案例研究峰会

大数据环境下实现一个通用推荐引擎的实践

邓雄

58同城数据智能部 总监

中科院大学工信学院大数据方向 特聘专家委员



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com



全球软件案例研究峰会

关于我

- **9年数据挖掘相关研究研发经验**
- **58同城数据智能应用部总监**
- **中科院大学工信学院大数据方向专家委员会特聘委员**
- **曾担任人人网应用研究中心、清华联合实验室负责人**
- **曾研发百度商务搜索部凤巢广告**
- **英国帝国理工 数据挖掘 PhD**
- **受邀演讲：**
 - ✓ IBM Ireland Research Center (In English) , 2010
 - ✓ 中国系统架构师大会, 2013.9
 - ✓ 杭州阿里技术分享, 2013.10
 - ✓ 中国软件技术大会, 2013.12
 - ✓ CITC全球互联网技术大会, 2013.12.5
 - ✓ Top100 Summit全球软件案例研究峰会, 2013
 - ✓ 58同城大数据力量系列讲座, 2014

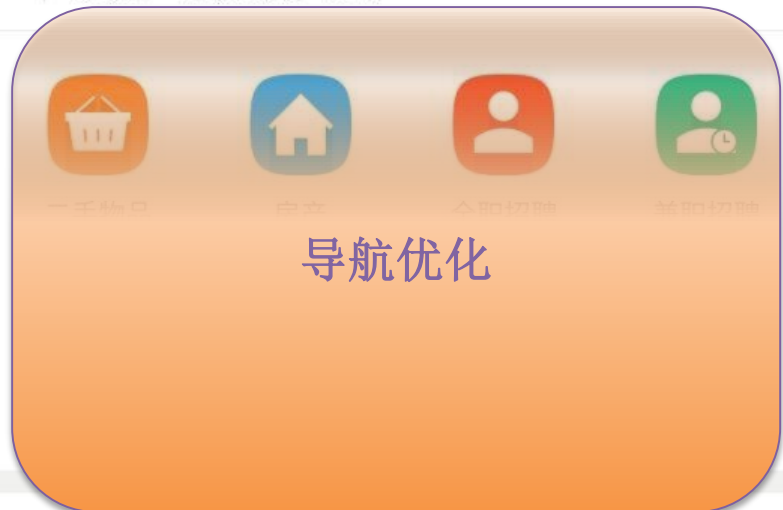


TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com



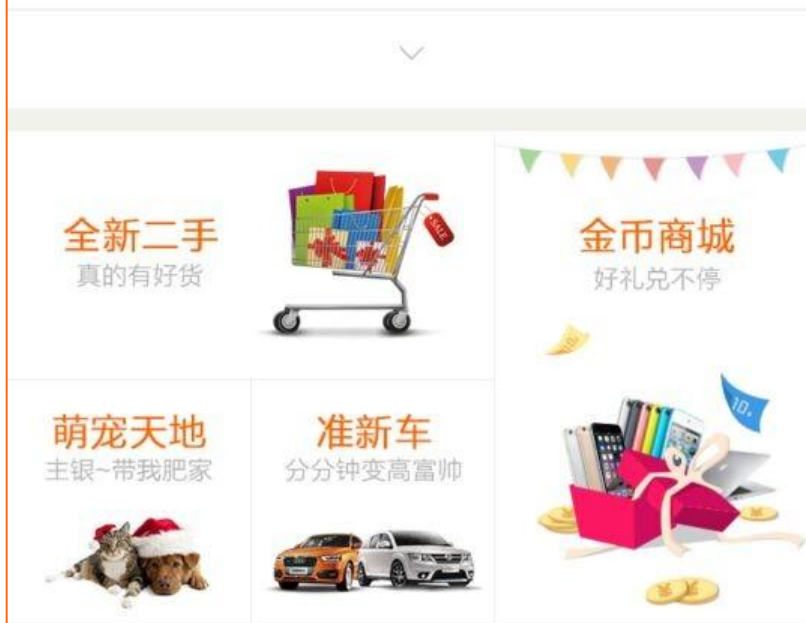
足迹：地铁线路 导购



58自营服务 100%服务专业 100%价格实惠 100%赔付保障



执门服务



全职招聘

发布信息

请输入类别或关键字

热门专区

为我优选

五险一金

周末双休

附近工作

包吃住专区

应届生专区

热门职位

标签推荐

导购

促销/导...

薪资

福利

更多

导购员+吃住+底薪+提成

朝阳-亚运村小营 北京歌斯堡酒业有限公司

3000-5000元/月 今天

总部急聘销售+有无经验

朝阳-媒体村 北京我爱我家房地产经纪有限公司

5000-8000元/月 今天

智能排序

北京大屯朝阳 佳园房地产经纪有限公司

全城

附近

导购

区域

促销/导...

薪资

更多

CK全北京高薪急聘导购

朝阳 美商华尔纳商贸（上海）有限公司

3000-5000元/月 今天

中国移动营业厅手机营业员

北京 北京中复电讯设备有限责任公司

5000-8000元/月 今天

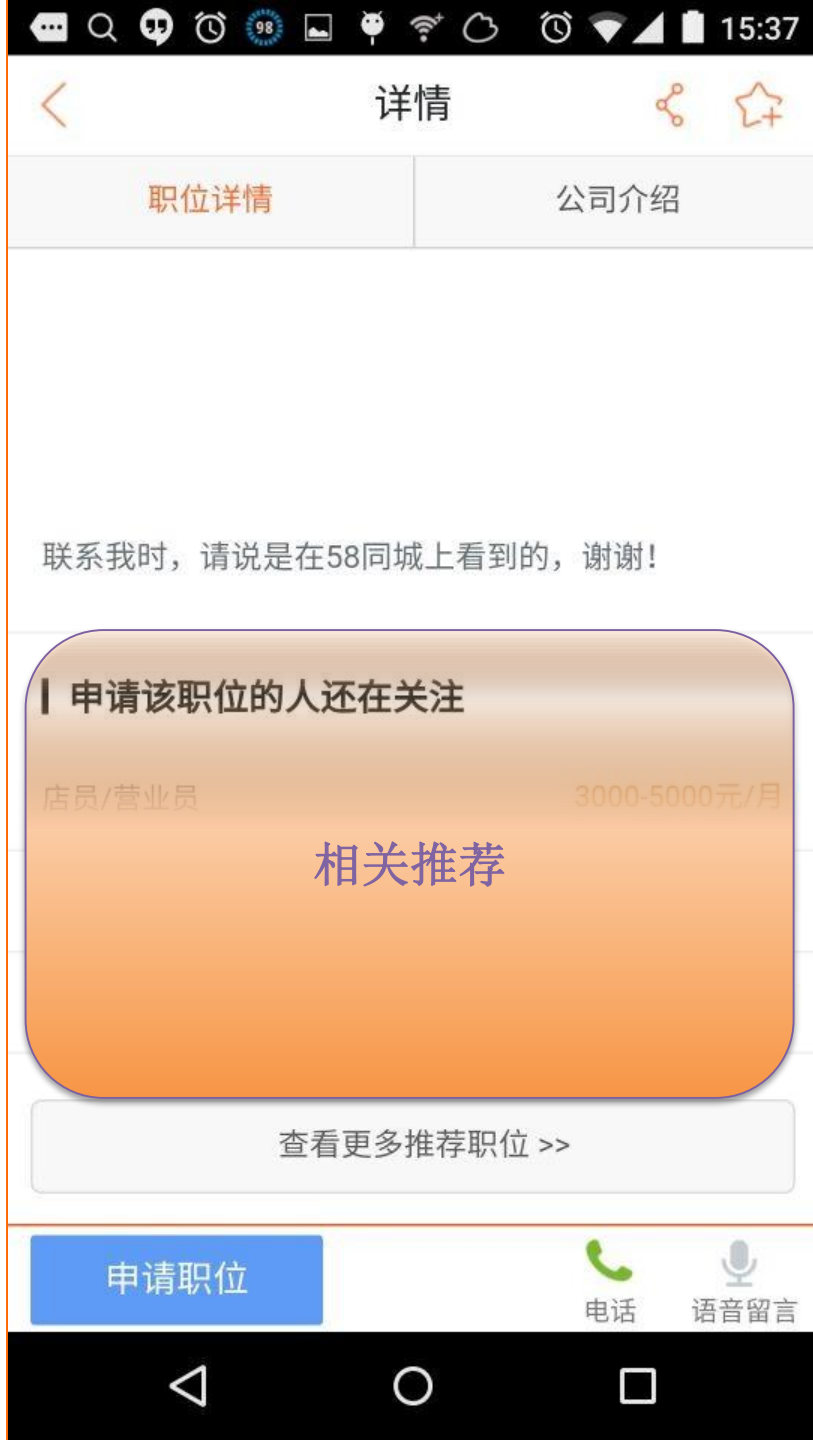
计算广告

北京 北京中复电讯设备有限责任公司

5000-8000元/月 今天

全城

附近





全球软件案例研究峰会

你能了解到什么？

- 推荐引擎解决的问题
- 推荐引擎历史
- 推荐引擎基本原理
- 通用基础架构
- 推荐引擎相关算法



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com



全球软件案例研究峰会

大数据背景下的推荐引擎主要挑战？

- 信息爆炸、信息过载

- 1分钟互联网产生多少数据？

- 48小时新视频@Youtube
 - 2,000,000次搜索请求@Google
 - 684,478分享消息@Facebook
 - 100,000条tweets@Twitter
 - 3600张照片@Instagram



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

大数据背景下的推荐引擎主要挑战？

• 智能化、移动化、人性化

- Web智能：搜索网站、购物网站、社交网站、计算广告
- App智能（2014年十大APP）
 - 移动O2O、支付
 - 移动交友、通讯
 - 移动新闻、视频分享
 - 移动安全
- 智能硬件
 - 智能家居：智能电视、智能路由、智能冰箱、智能安防
 - 移动智能设备：可穿戴设备、智能车载设备





全球软件案例研究峰会

大数据背景下的推荐引擎主要挑战？

*We are moving from an **Information Age**
to the **Recommendation Age**.*

– “*The Long Tail*” by Chris Anderson



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com



全球软件案例研究峰会

- **推荐系统**：发现用户偏好，给用户主动推荐符合其意图的信息
 - 好友推荐，商品推荐，网络日志推荐，视频推荐，App推荐，广告推荐
 - Amazon, Facebook , Google, Netflix, Youtube, Apple...



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

• “推荐引擎是未来互联网的发动机”

– Netflix: “让你喜欢的电影“跳”出来”

- 1997，成立，主营DVD租赁，O2O
 - ① 片源分类、汇总整理
 - ② 制定价格、组建渠道、开展促销
- 1999，订阅服务：Cinemath推荐引擎
 - ① 点评、电影特征、环境影响
- 2006，Netflix百万美金推荐大赛
- 2010，年收入20亿美金，注册用户1730万，付费用户超过500万，点评数据30亿条，售出10亿份DVD
- 2011，在线电影销售占全美45%，超过Apple
- 2013，基于大数据投拍电视剧：《纸牌屋》





全球软件案例研究峰会

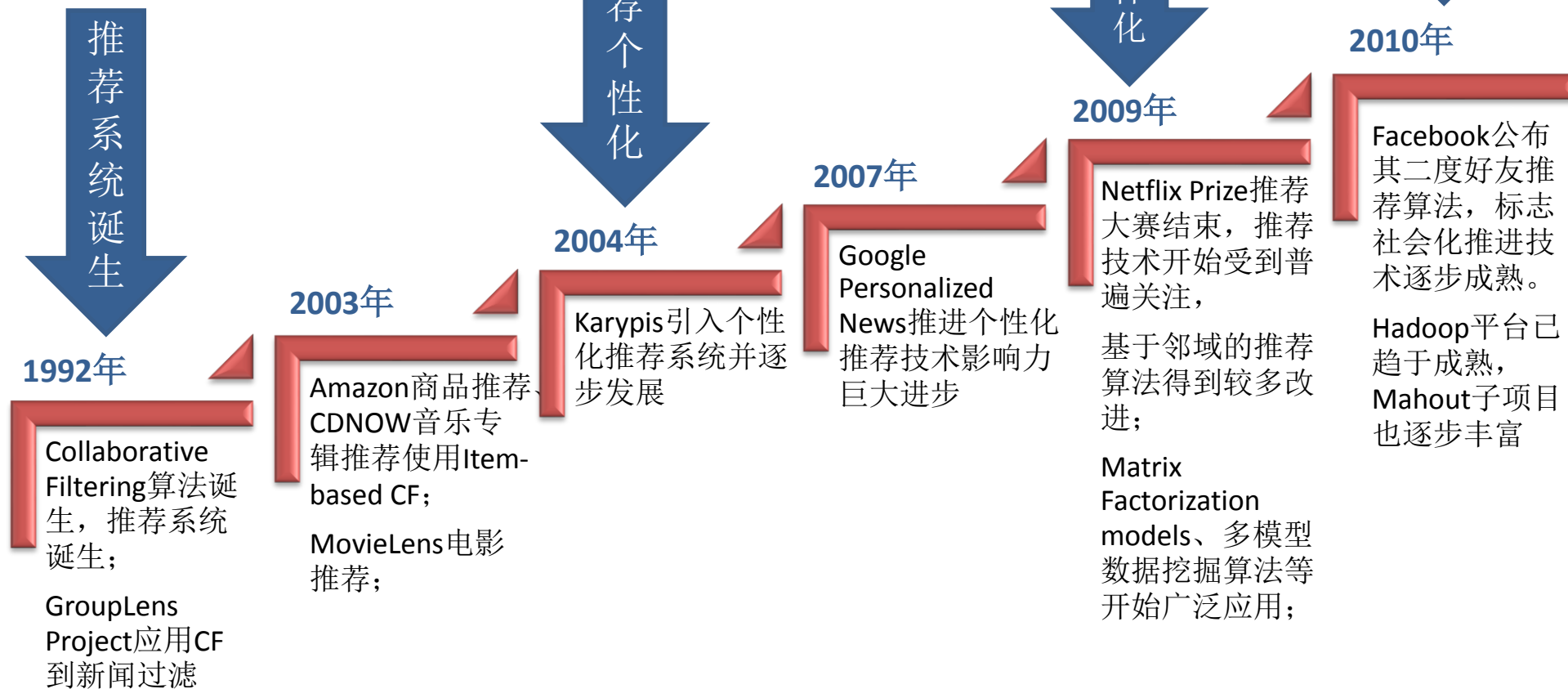
你能了解到什么？

- 推荐引擎解决的问题
- 推荐引擎历史
- 推荐引擎基本原理
- 通用基础架构
- 推荐引擎相关算法



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com





全球软件案例研究峰会

你能了解到什么？

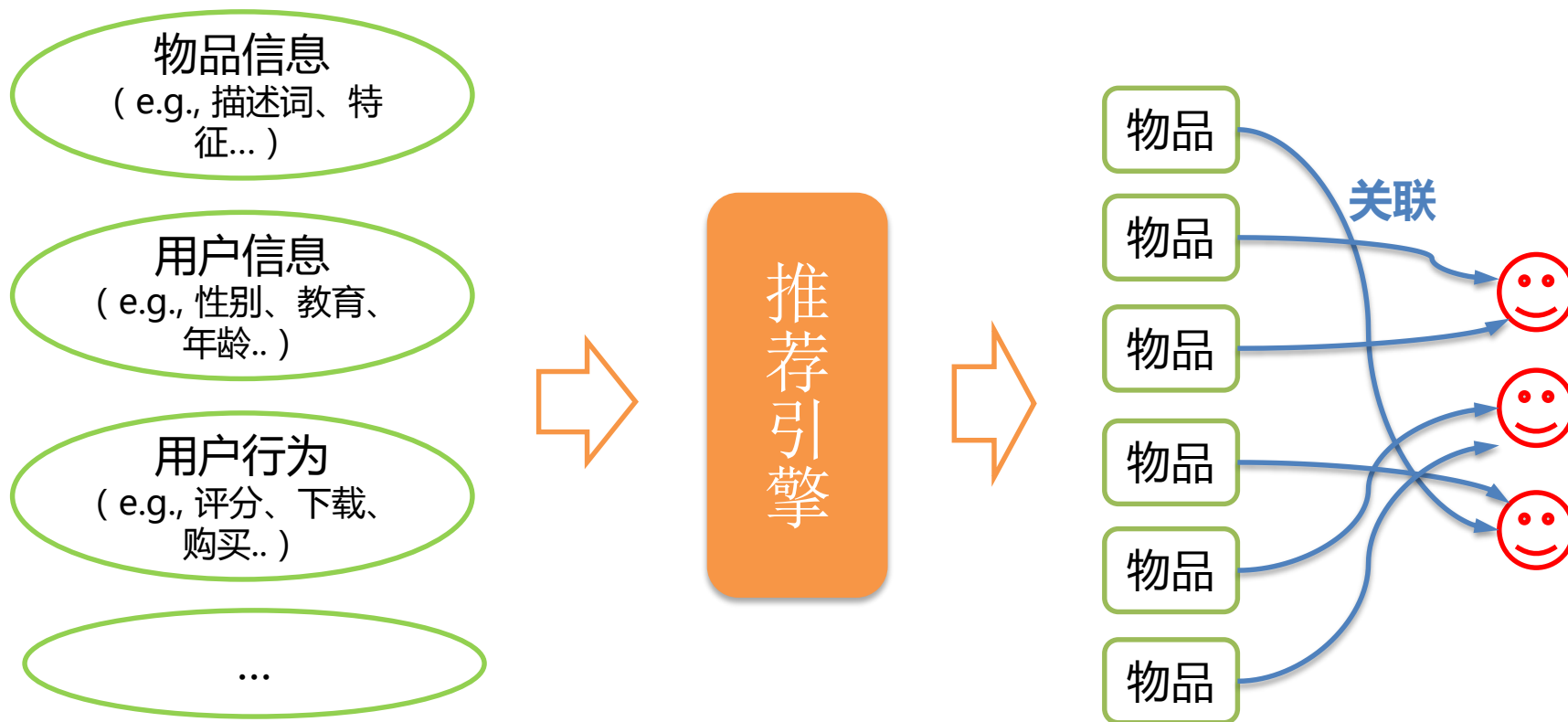
- 推荐引擎解决的问题
- 推荐引擎历史
- **推荐引擎基本原理**
- 通用基础架构
- 推荐引擎相关算法



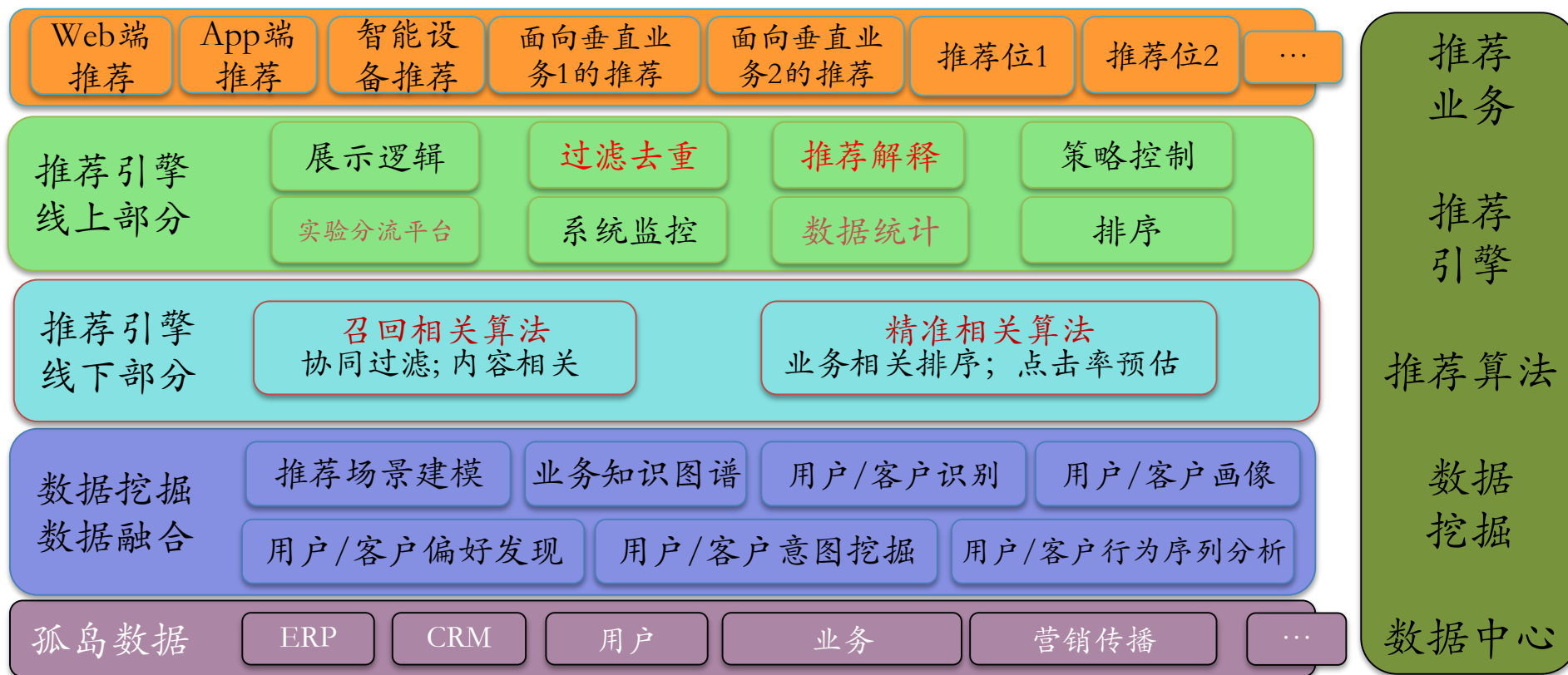
TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

• 推荐引擎通用工作原理



通用推荐引擎分层体系架构





全球软件案例研究峰会

你能了解到什么？

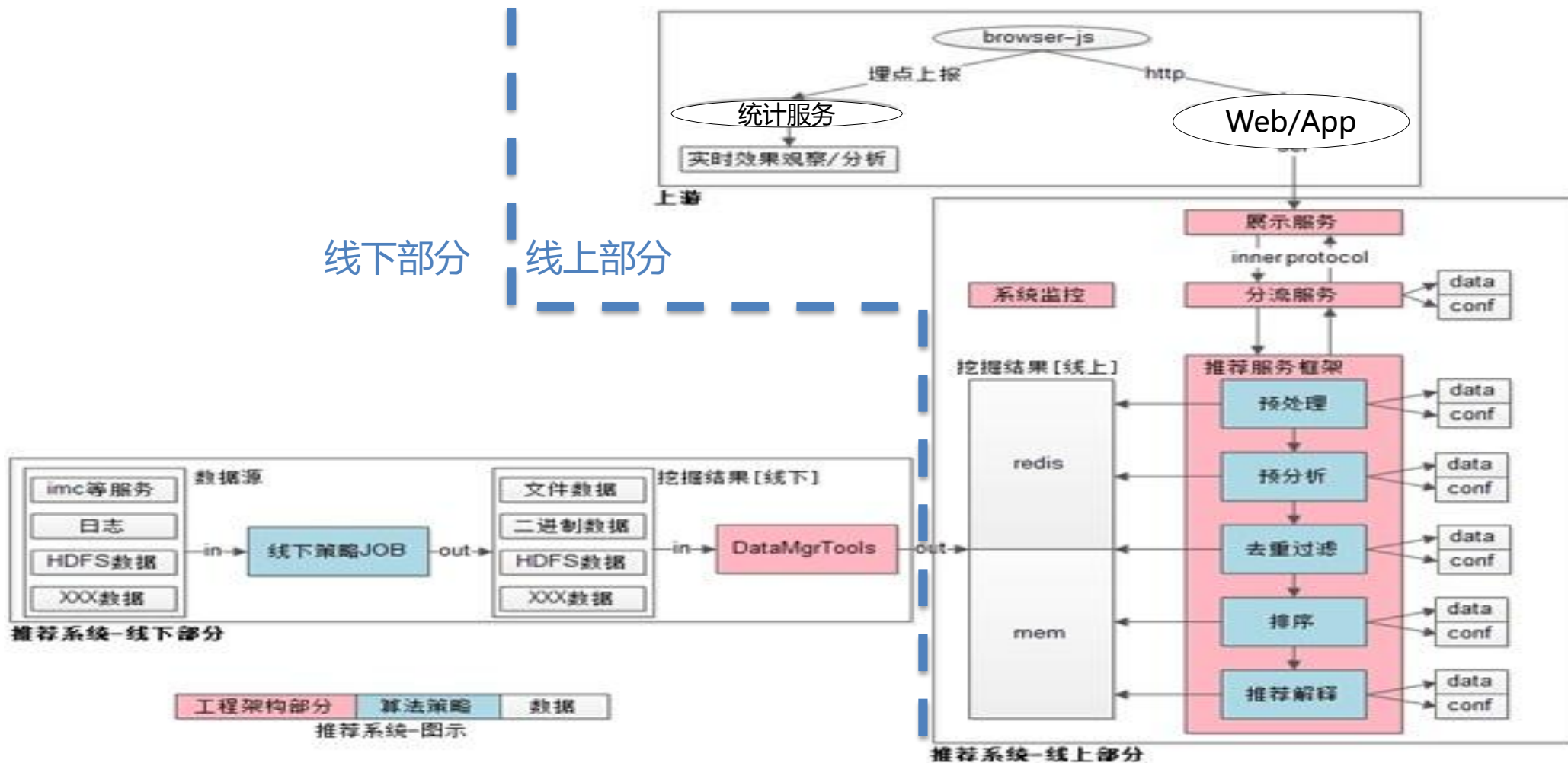
- 推荐引擎解决的问题
- 推荐引擎历史
- 推荐引擎基本原理
- **通用基础架构**
- 推荐引擎相关算法



TOP 100 CASE STUDIES
OF THE YEAR

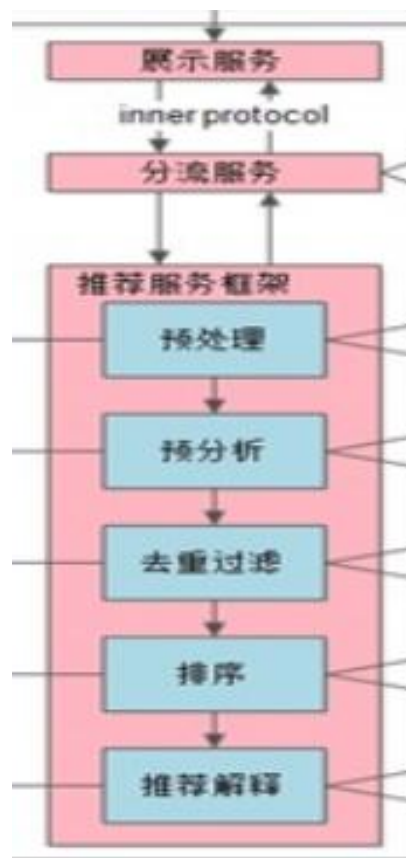
www.top100summit.com

通用推荐引擎基础架构

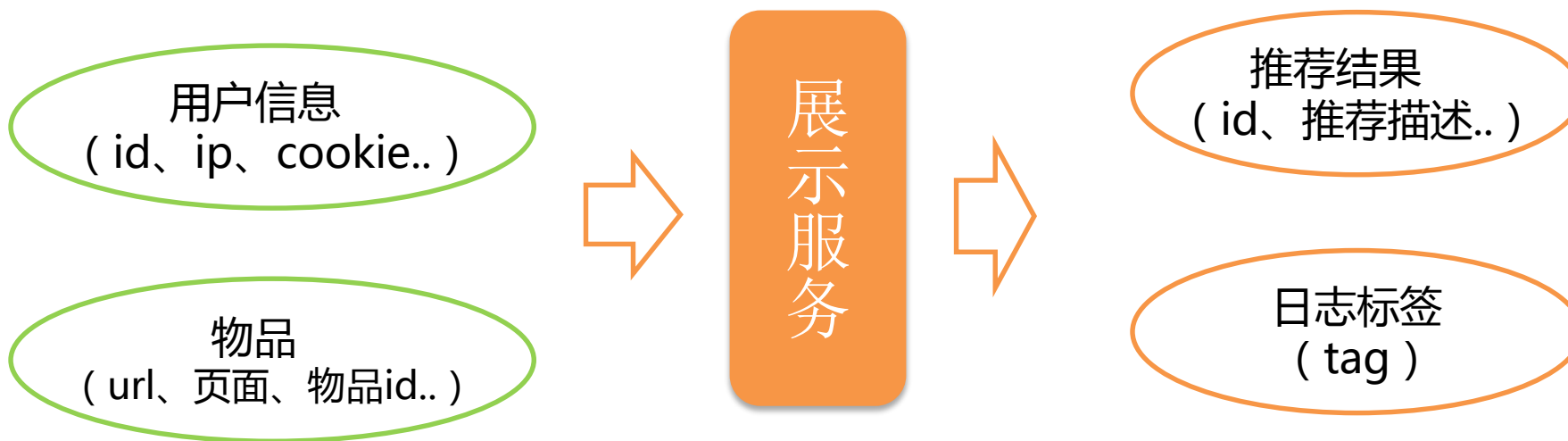


通用架构关键模块

- 线上架构部分
 - ① 统一展示逻辑
 - ② 实验分流平台
 - ③ 推荐内核

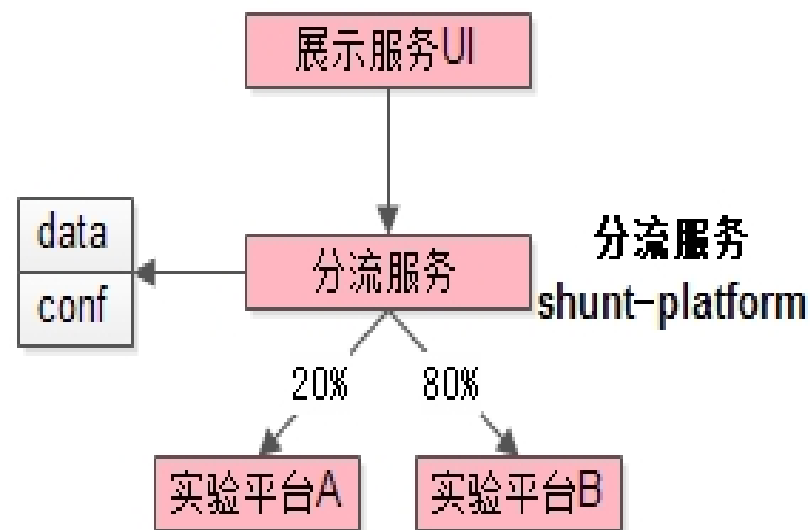


线上部分：展示服务



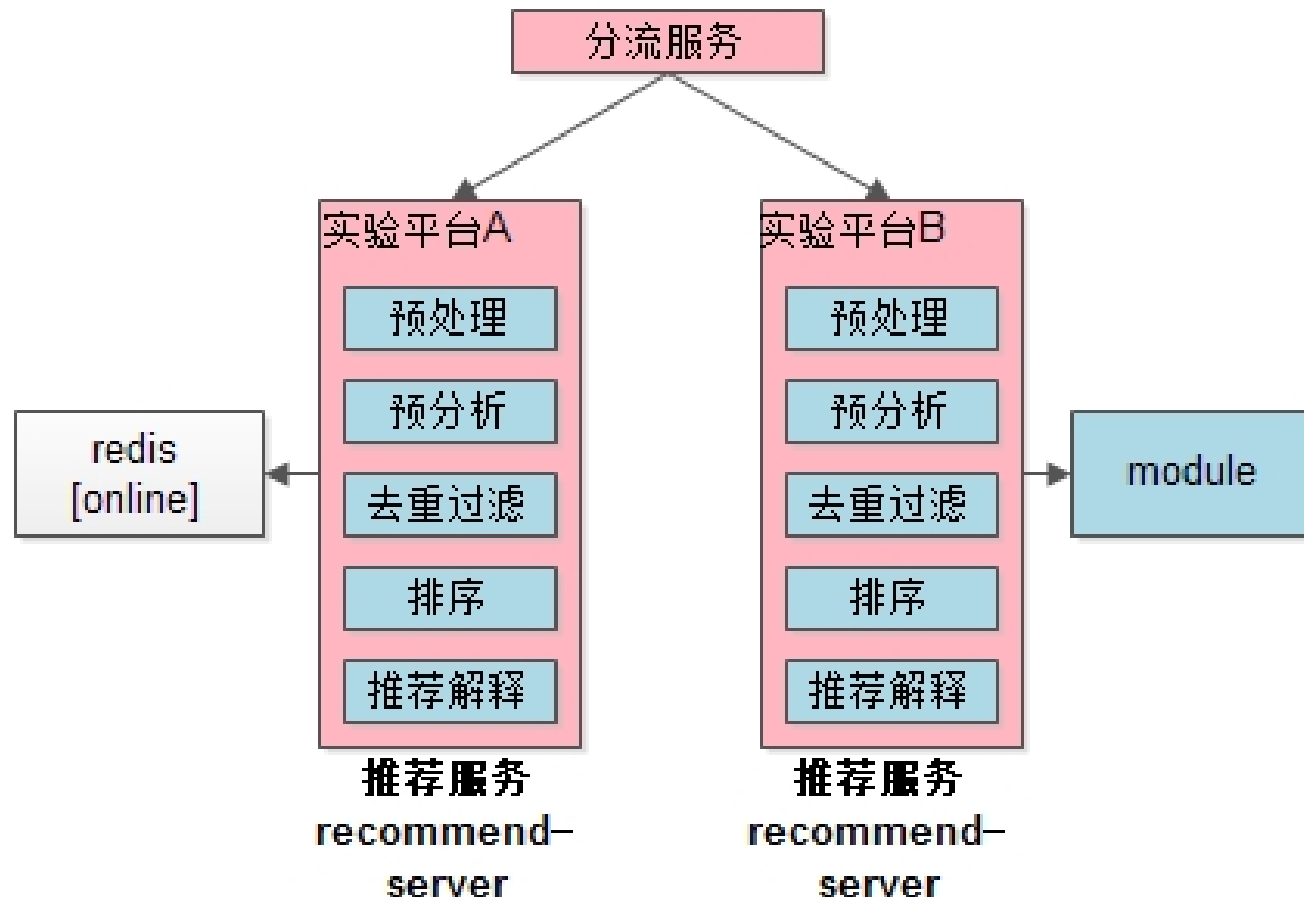
线上部分：实验分流平台

- ① 根据配置规则决定分流：ip=xxx && area == Guangzhou；
- ② 黑白名单分流：if(uid in whitelist)；
- ③ random分流



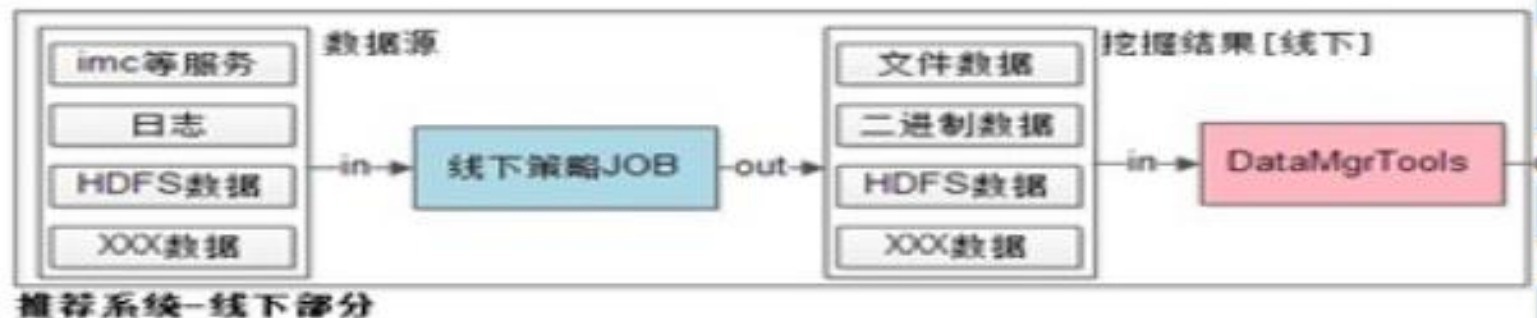
线上部分：推荐内核

- ① 结果召回
- ② 去重过滤
- ③ 排序
- ④ 推荐解释



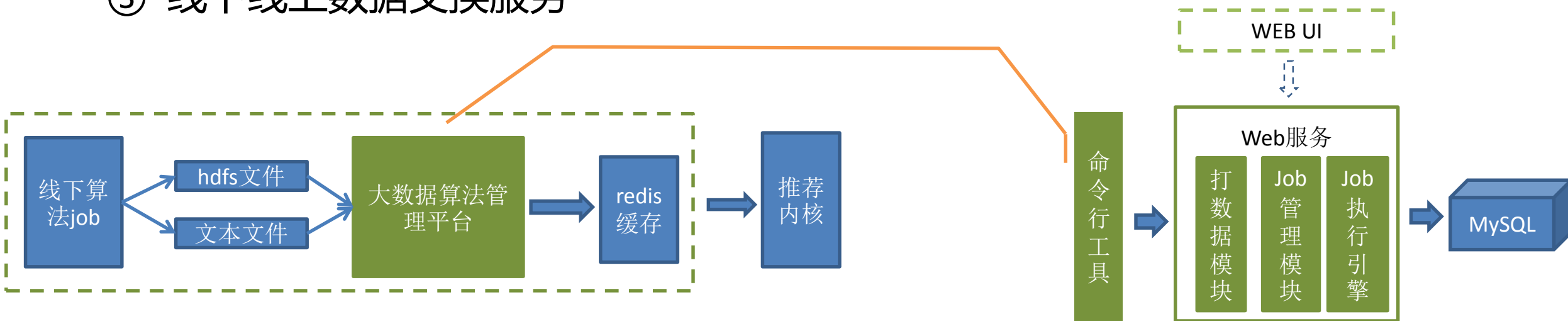
通用架构关键模块

- 线上架构部分
 - ① 统一展示逻辑
 - ② 实验分流平台
 - ③ 推荐内核
- (半)线下架构部分
 - ④ 实时数据统计分析平台
 - ⑤ 数据挖掘和推荐算法管理平台

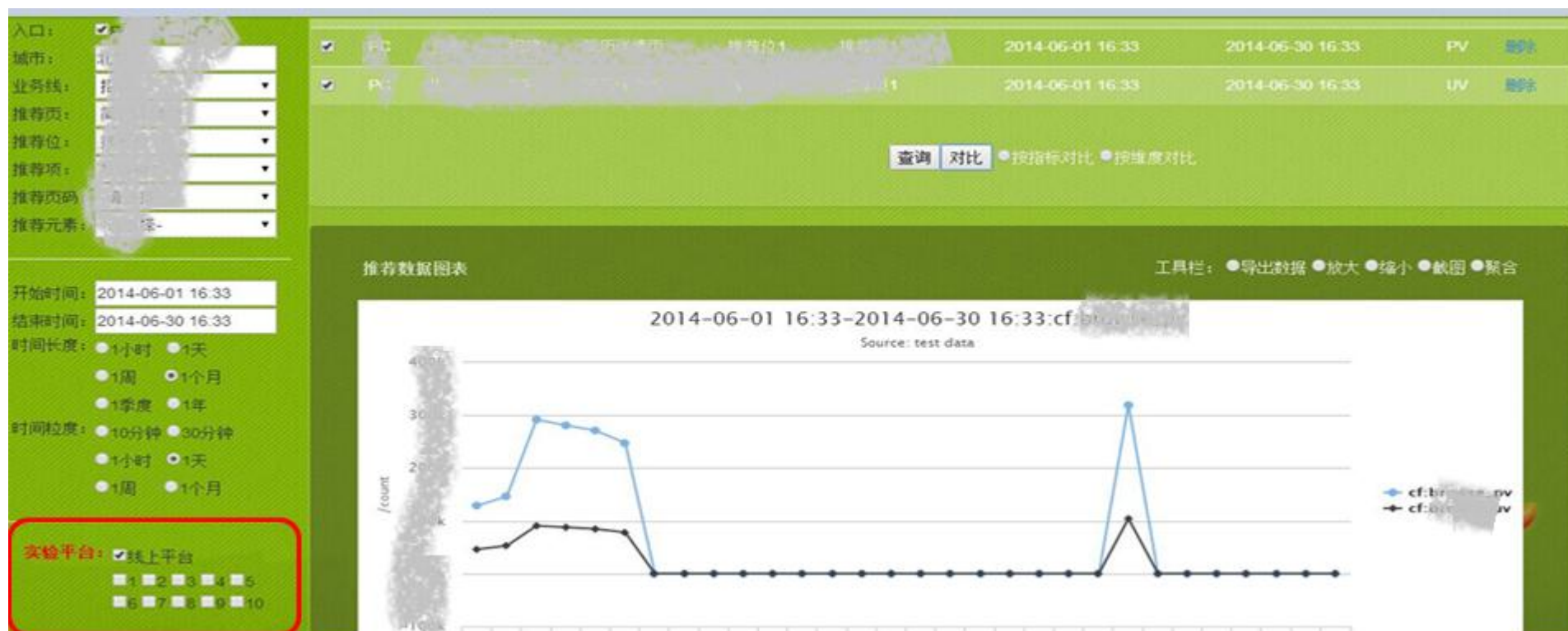


(半) 线下部分：算法管理平台

- ① 大数据清洗、收集、转化
- ② 线下挖掘算法的输入数据、中间数据、输出数据管理
- ③ 线下线上数据交换服务



(半) 线下部分：实时业务效果分析平台





全球软件案例研究峰会

通用架构关键模块

- 线上架构部分
 - ① 统一展示逻辑
 - ② 实验分流平台
 - ③ 推荐内核
- (半)线下架构部分
 - ④ 实时数据统计分析平台
 - ⑤ 数据挖掘和推荐算法管理平台
- 监控系统

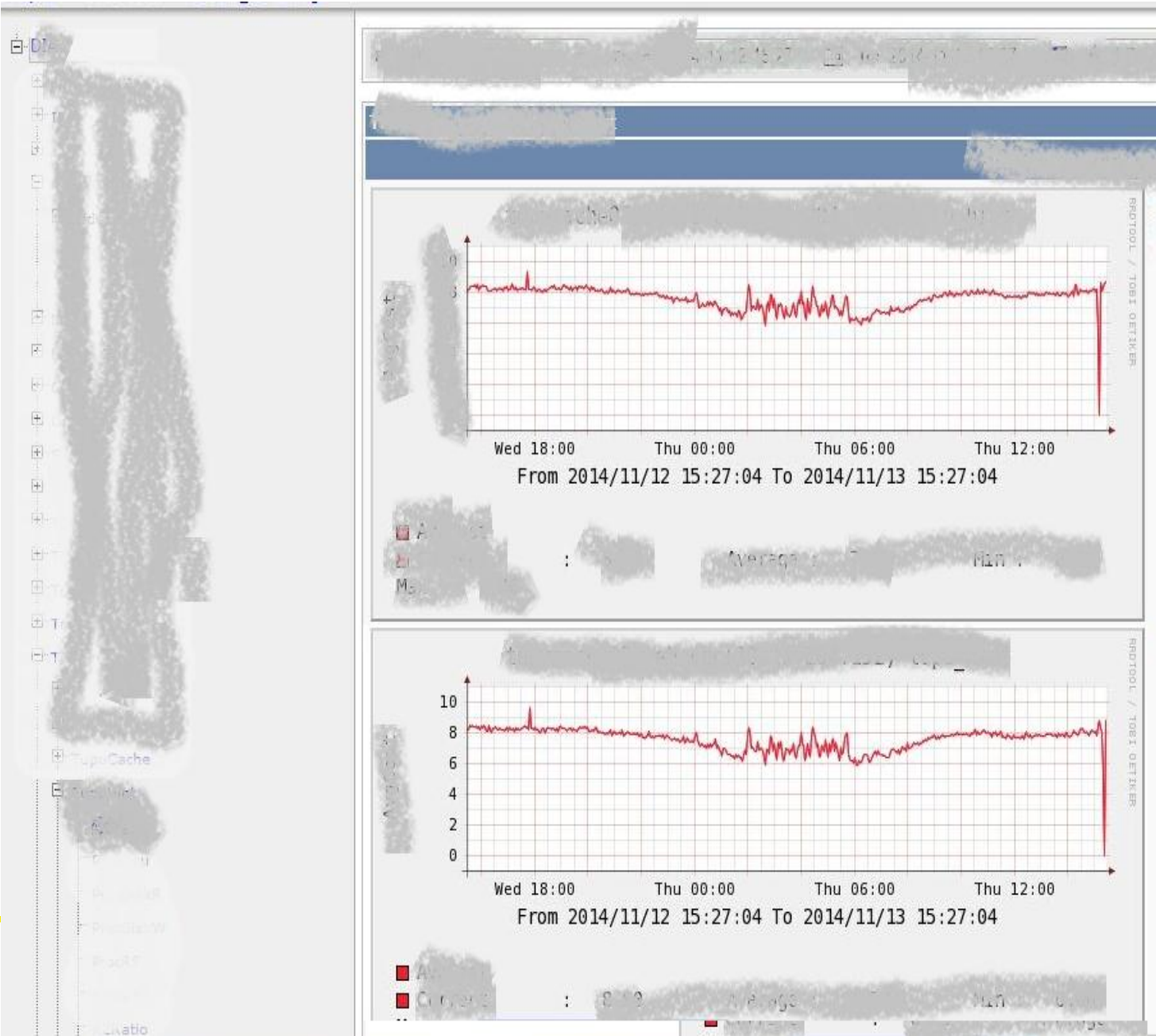


TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

系统监控

- ① 硬件级别
- ② 系统级别
- ③ 接口/服务级别
- ④ 业务数据监控





全球软件案例研究峰会

你能了解到什么？

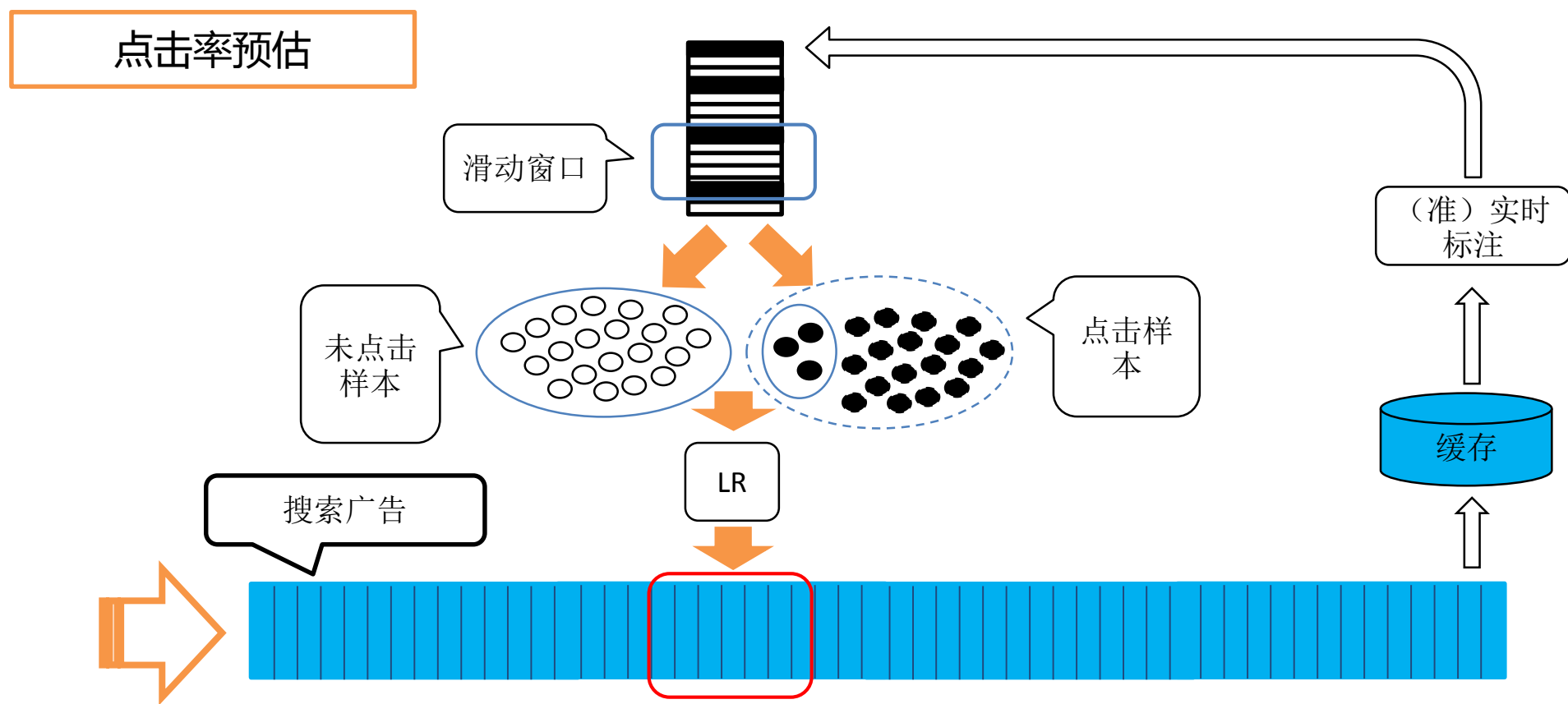
- 推荐引擎解决的问题
- 推荐引擎历史
- 推荐引擎基本原理
- 通用基础架构
- 推荐引擎相关算法



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

Online Learning的数据特点和一般流程





全球软件案例研究峰会

核心推荐算法相关库

- 全局唯一用户识别GUID：不能标识用户（群）的具体行为



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

电影 > 大陆 > 喜剧/剧情/优酷出品

老男孩猛龙过江：老男孩猛龙过江

淘宝网
Taobao.com



¥145.00

天猫

YOUKU
优酷



公映许可证
电审故字[2014]第210号

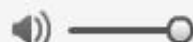
国家新闻出版广电总局电影局

FILM BUREAU, STATE ADMINISTRATION OF PRESS, PUBLICATION, RADIO, FILM AND TRLEVISION

<< 播放列表



00:05 / 114:06



高清



电影 > 大陆 > 喜剧/剧情/优酷出品

老男孩猛龙过江：老男孩猛龙过江

淘宝网
Taobao.com



¥145.00

天猫

YOUKU
优酷



登录



累积
观看时长



提升
观看等级



超越
小伙伴

快速登录



QQ



微博



支付宝

使用优酷帐号

登 录

注 册

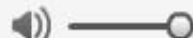
FILM BUREAU, STATE

LM AND TRLEVISION

<< 播放列表



00:05 / 114:06

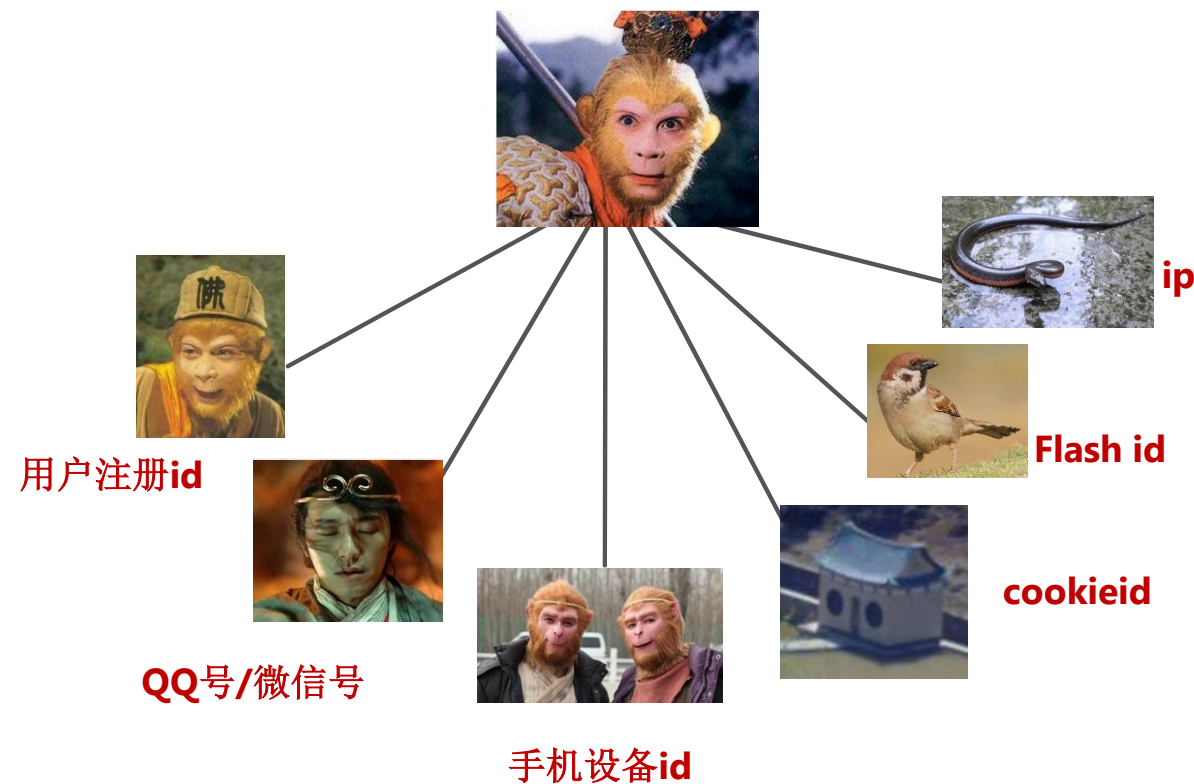


高清



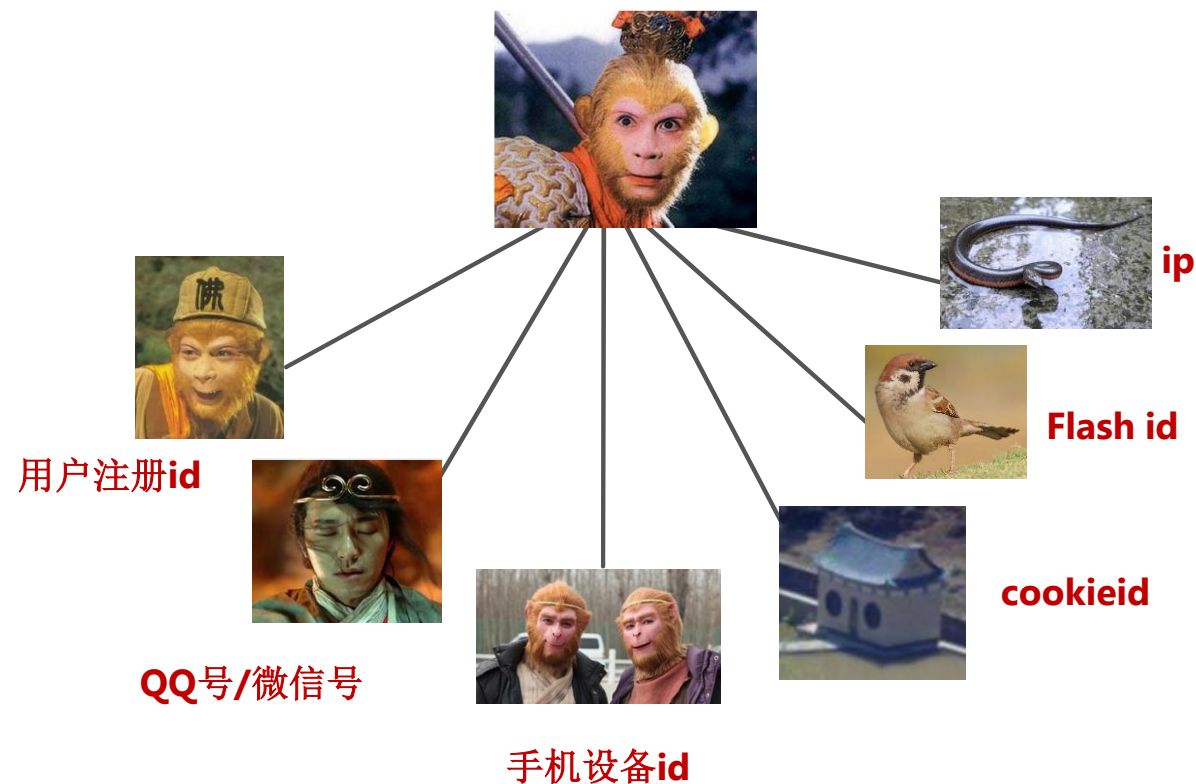
核心算法：全局唯一用户识别GUID

- 问题：不能标识用户（群）的具体行为
 - ① 大部分浏览型应用的用户持续未登录浏览
 - ② 多次未登录浏览后再登录
 - ③ PC、M、App多入口同时登录



核心算法：全局唯一用户识别GUID

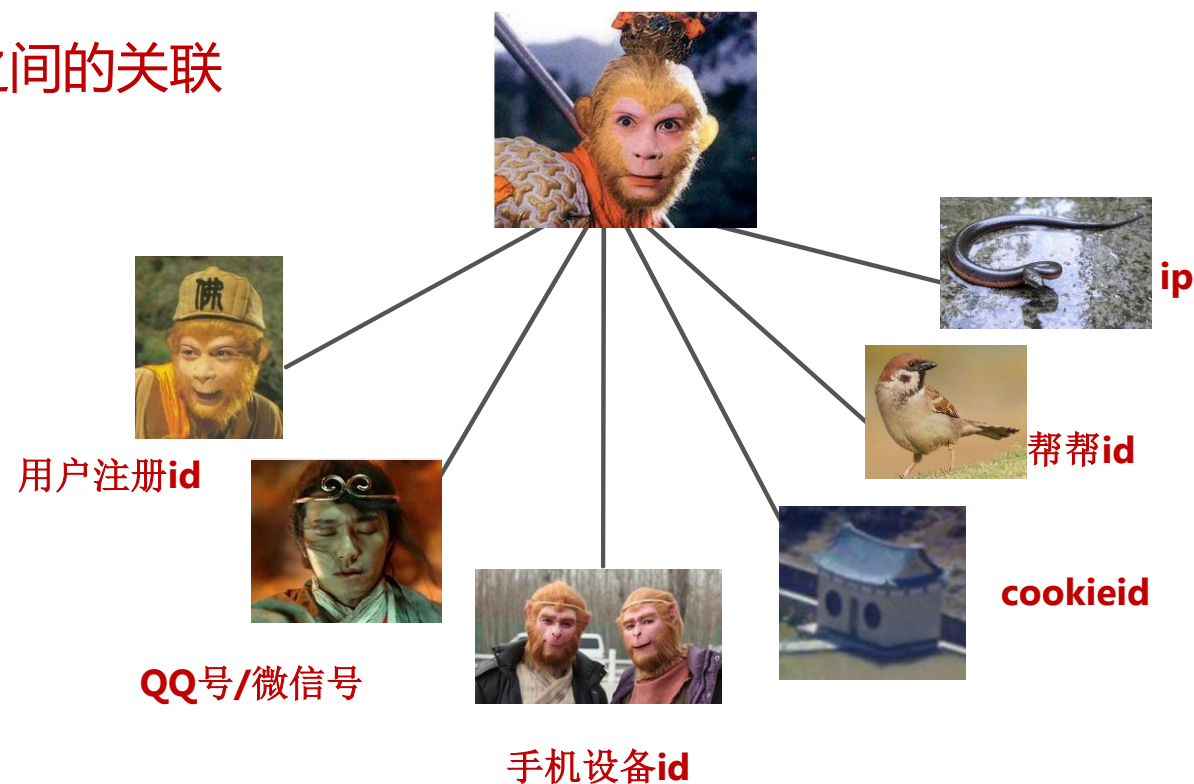
- 影响：**大数据价值难挖掘**
 - ① 流量：无法实现精细化流量管理；
 - ② 收入：广告精准定向难以实现，收入效率难以大幅提升；
 - ③ 市场运营：难以精细理解自身优势目标客户特点，营销运营难以精准化和随势而变；



核心算法：全局唯一用户识别GUID

- 方案：分析用户每次访问特征信息，建立特征之间的关联

- ① 硬关联：cookie、flash-id、imei、ip
userid、QQ号/微信号
– 利用登录行为、手机使用行为管理
- ② 软关联：动态行为聚类
– 从行为轨迹和点击内容上判别与历史用户关联





全球软件案例研究峰会

核心推荐算法相关库

- 用户及业务画像：用户定向、业务价值最大化、营销指导



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

核心算法：用户及业务画像

- 意义及价值：用户定向、业务价值最大化、营销指导

- ① 基本属性特征
- ② “衣食住行” 相关兴趣特征
- ③ 业务相关商业价值特征



上海; PC;
10:00,12:00, 17:00

科技迷 关注时事
娱乐八卦 关注女性 玩游戏
爱旅游 体育迷 研究
星座 博客控 **F1** 音乐
迷 财经高手 关注教育 汽
车迷 **高尔夫** 买彩票 哈
韩 田径 **读书狂** 军事迷
育儿 文化 重视健康 **足球**



青岛, 济南; PC, 9:00, 15:00

仁和可立克 **科技迷** 博客控 爱
看视频 娱乐八卦 美容 汽车迷 财经高手

北京; PC, Iphone; 9:00, 12:00, 21:00

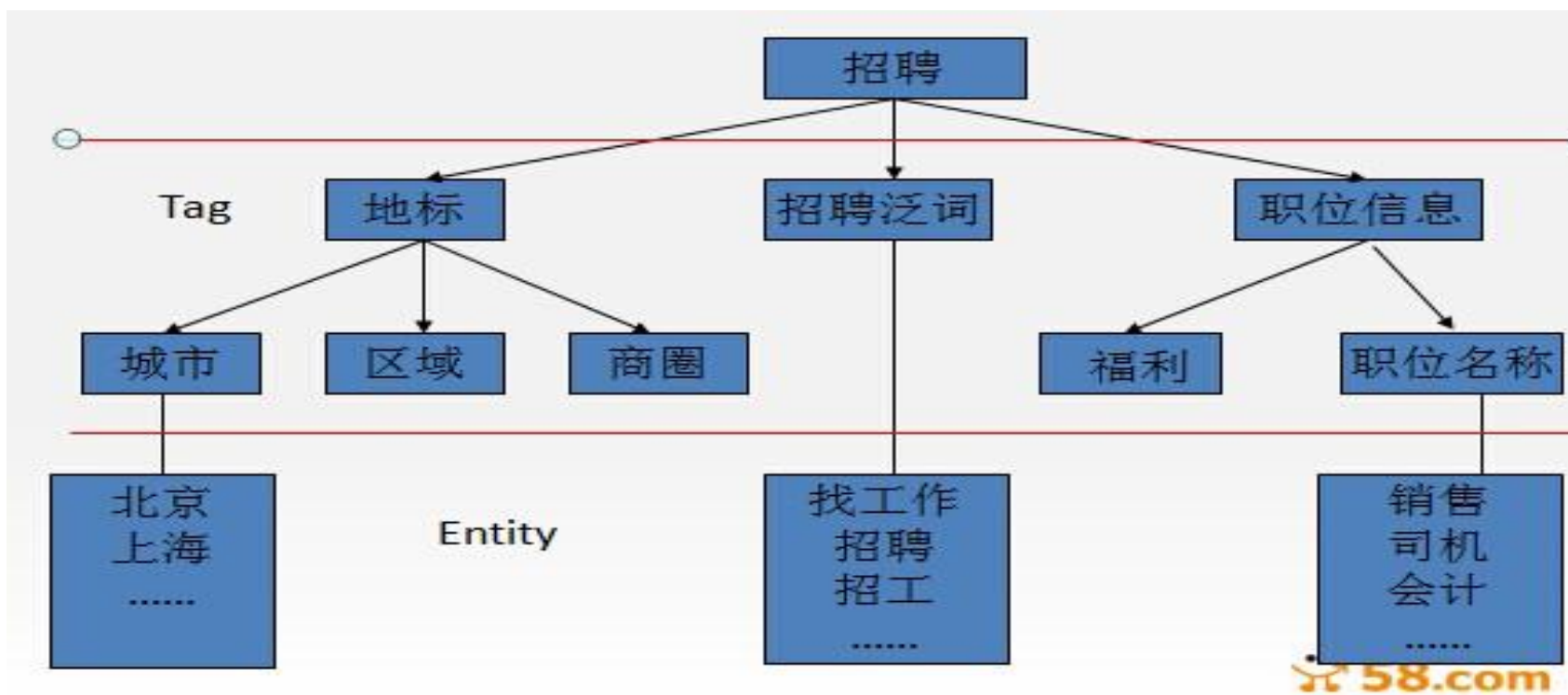
雪花啤酒 田径 **汽车迷** 游泳 房产
奔驰车 羽毛球 **体育迷** **三星**
玩游戏

核心算法：用户及业务画像

- 影响：**大数据价值难挖掘**
 - ① 流量：无法实现精细化流量管理；
 - ② 收入：广告精准定向难以实现，收入效率难以大幅提升；
 - ③ 市场运营：难以精细理解自身优势目标客户特点，营销运营难以精准化和随势而变；



核心算法：用户及业务画像





全球软件案例研究峰会

核心算法：用户及业务画像

The screenshot displays a real estate website interface with two columns of property listings. Each listing includes a small image, a title, a description, and a price. The listings are for various properties in different areas, such as '天通苑' (Tiantongyuan) and '北三区' (North Area 3). The prices range from 3800 to 5500 yuan per month. The website URL is visible at the top of the browser window.

Property ID	Title	Description	Price (Yuan/month)
11图	直租优质房源天通苑西三区 3室2厅146平米 精装修	采光好 出行方便 精装修 家具家电齐全	5500
11图	低价精装三居 天通苑北三区 3室2厅145平米	住家子 家具家电齐全 只能自住 房东非常爱干净	4000
8图	【链家100%真房源】西三区精装好房随时能看精装家具	大四居 拎包入住 地铁房	4100
9图	免佣房源 北三区 4室3厅200平米 中等装修	免佣房源 有钥匙 随时看房 可以租两年 不涨价	5400
6图	天通苑北三区 3室2厅145平米 有钥匙 随时看房		3800



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com



全球软件案例研究峰会

核心推荐算法相关库

- 实时CTR预估：决定结果排序的最重要依据



TOP 100 CASE STUDIES
OF THE YEAR



www.top100summit.com

核心算法：实时CTR预估

- 意义价值：决定结果排序的最重要依据

您可能感兴趣的简历 NEW		
普工或技工 	北京	
赵大众 男 31岁 无经验	2014-11-20	
普工或技工	北京	
刘蕊 女 23岁 无经验	2014-11-19	
普工或技工 	北京	
张占龙 男 36岁 3-5年经验	2014-11-20	
普工或技工	北京	
柴萧强 男 20岁 无经验	2014-11-19	

北京二手手机

个人 商家 ☐  支持保障交易 ☐ 只看有图 ☐  帮帮在线





高价收售 IPHONE IPAD 笔记本等电子产品 [3图] 

上门取货电话 13651263420 价格合理，服务周到

商家推广 海淀 / 今天





思瑞凡老店2手应有尽有  1年 [24图] 

本店出售正品苹果以及三星九五成新的手机 支持先验货再付款 支

商家推广 海淀 - 北京大学 / 今天



实体店销售三星w2013.w999.w899  1年 [7图] 

中关村三星体验店出售w2013.999.899.s4.s3实

商家推广 海淀 - 中关村 / 今天

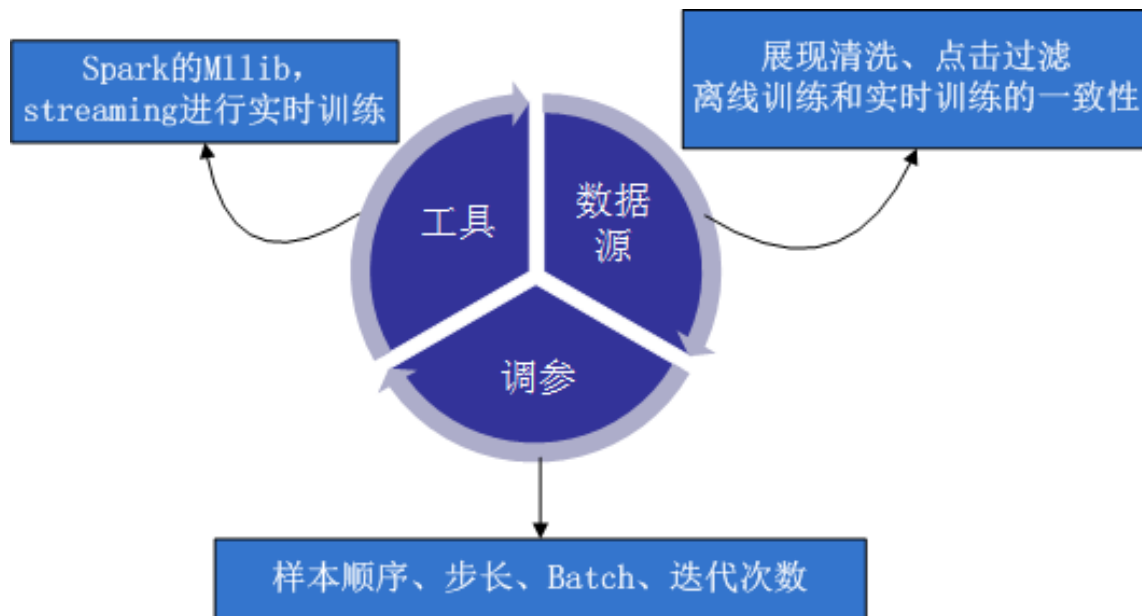


TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com

核心算法：实时CTR预估

- 方案：基于Spark Streaming的模型训练和使用





全球软件案例研究峰会

Thanks



TOP 100 CASE STUDIES
OF THE YEAR

www.top100summit.com