

# ArchSummit

## 全球架构师峰会（深圳）2014



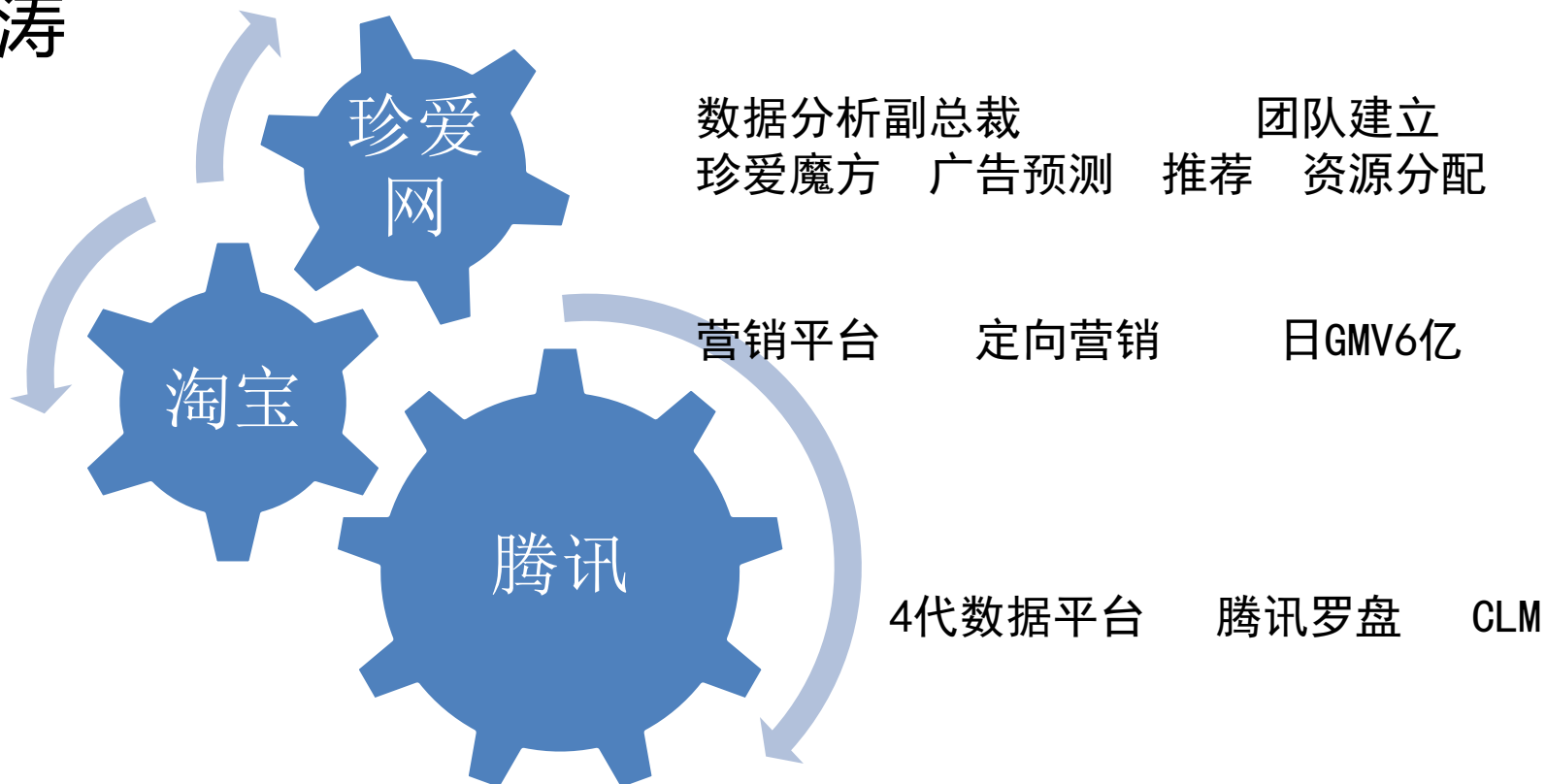
# 构建有业务价值的 数据分析系统

范成涛@珍爱网

# 自我介绍

---

范成涛



# 三件事情对我的触动

---

- 某数据牛人离开阿里
- 曾经的屌丝项目获得1.6亿的收入
- 我团队挖掘人员的故事

# 数据的几个误区

---

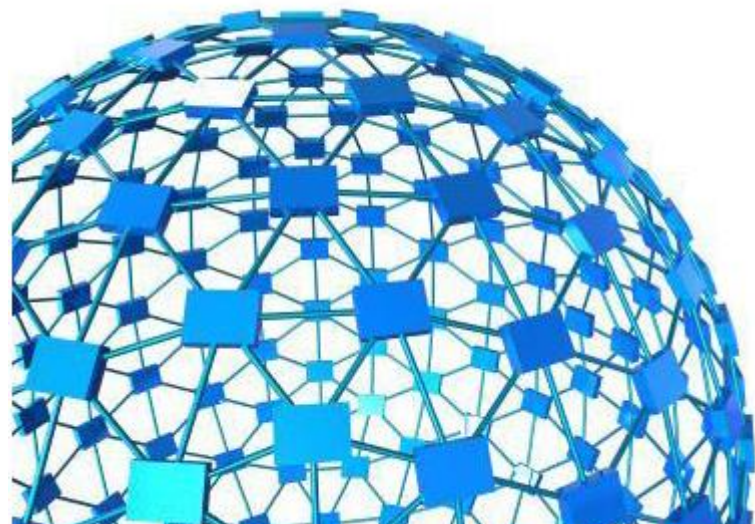
大 总是好的



快 总是有利的



大数据，技术 要牛B



# 数据的常用案例

---



# 数据工作的壁垒

---

业务人员

有强烈的业务提升需求

不知道数据能干啥

思维很受局限

技术人员

不了解业务

知道数据的技术实现

思维很受局限

# 数据的价值分层

---

了解业务现状

了解业务趋势

发现问题所在

认清用户

对接营销





# 常用分析方法经验分享

## 事前分析

- **如何预测各类数据**
  - 收入、用户数
- **如何建立考核指标**
  - 活动、功能
- **支持决策**
  - 科学决策，事实决策
- **精细化运营**
  - 准确触达，数据挖掘

## 事中分析

- **实时监控收入、用户**
  - 接触点效果反馈
  - 用户行为跟踪
- **实时查看流量**
  - 网络环境，运营环境
- **实时分析活动效果**
  - 根据效果调整策略

## 事后分析

- **产品有问题吗？哪里？**
  - 收入是否正常
  - 用户是否健康
  - 渠道是否安全
  - 流程是否顺畅
- **活动效果如何？**
- **功能效果如何？**
- **如何指导后续工作？**



# 常用分析方法经验分享

## 产品现状

- 来源、PV、UV、人数、次数
- 收入、arup、用户属性、活跃度

## 了解趋势

- 环比、同比、流动模型
- 增长率、留存率、流失率

## 发现问题

- 漏斗模型
- Ab test、调查问卷

## 认清用户

- 功能使用情况
- 热度分析

## 营销与推广

- 精准化投放
- 挽留、拉新

时间维度

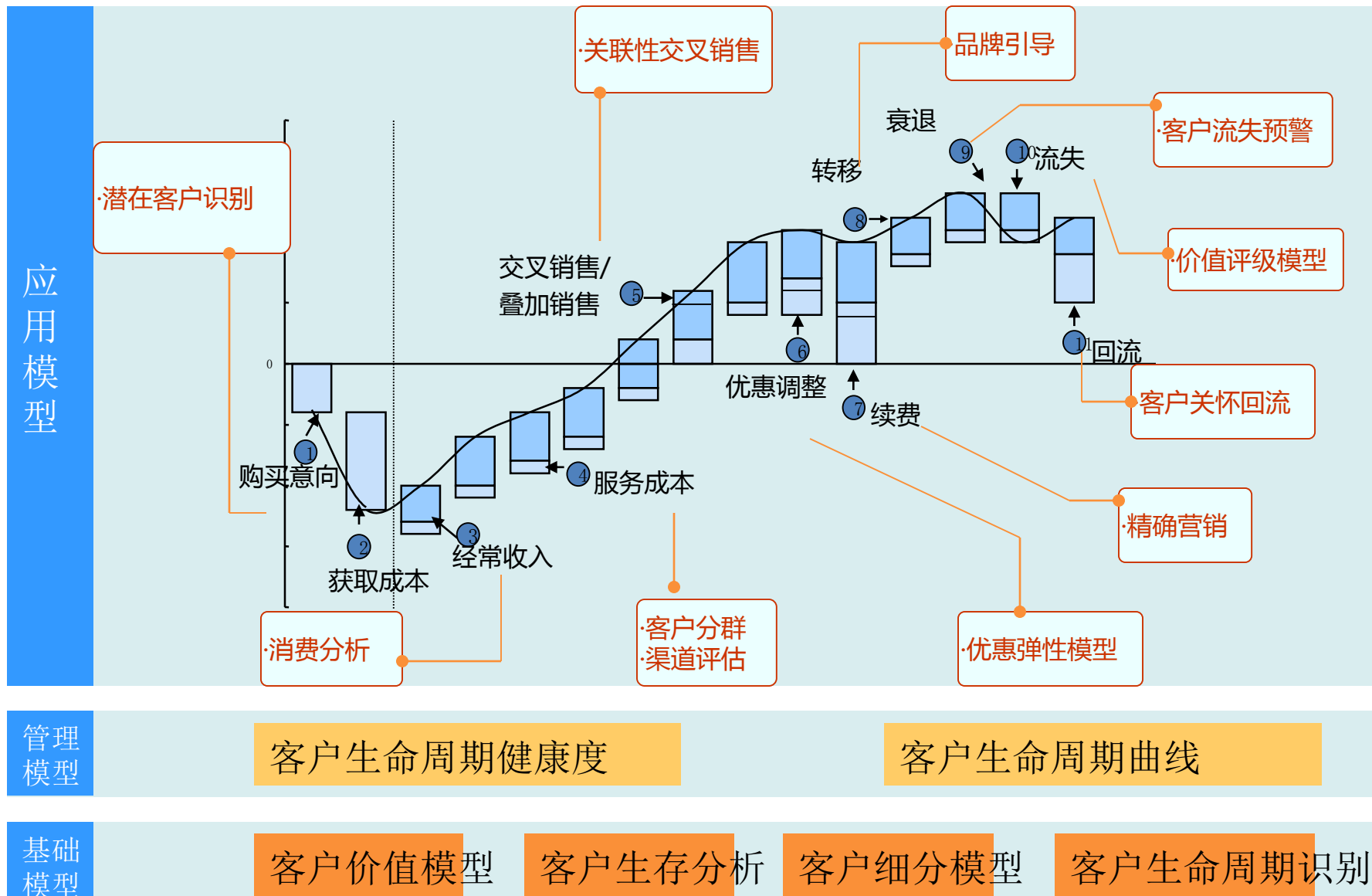
用户群维度

用户价值维度(等级等)

自然属性维度

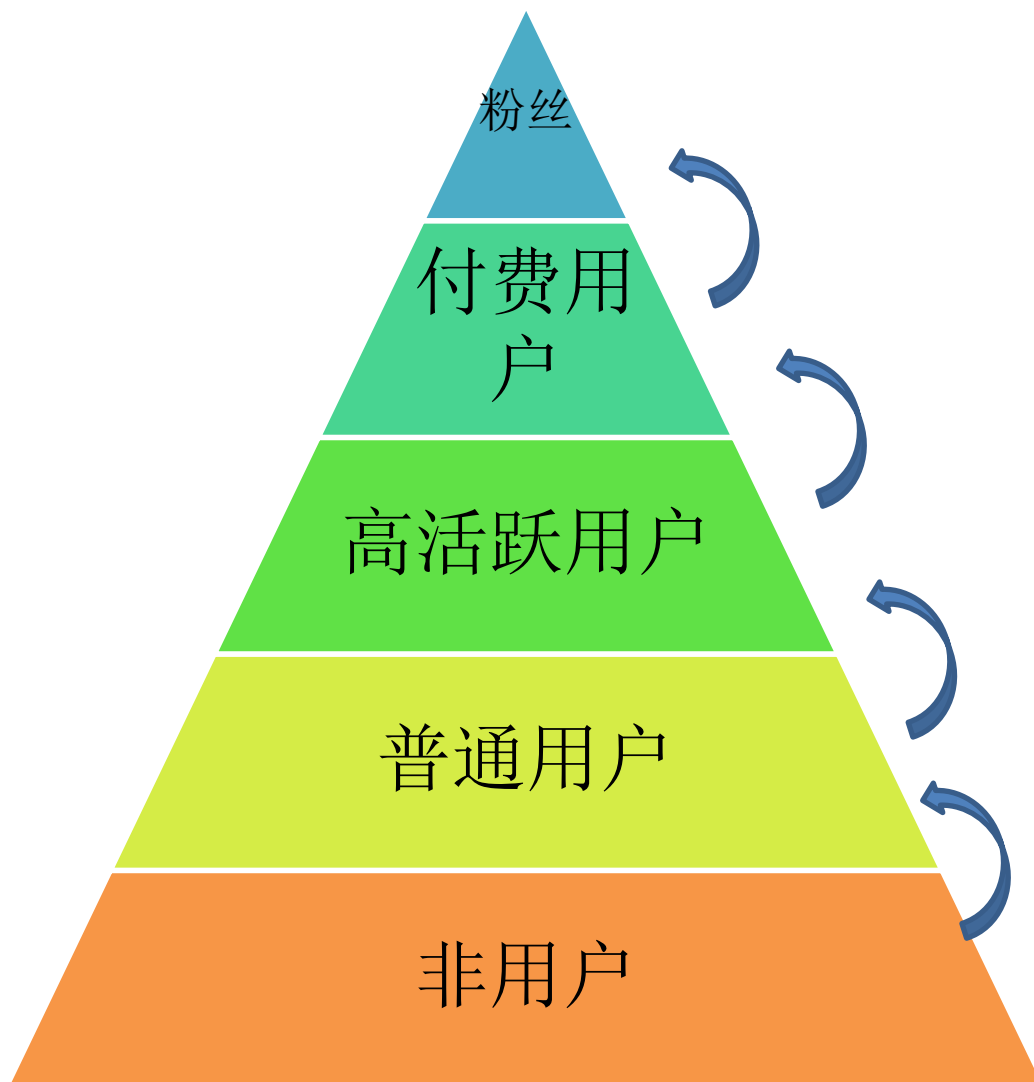
版本维度

# 生命周期管理示意图



# 用户生命周期周期管理的目标

---

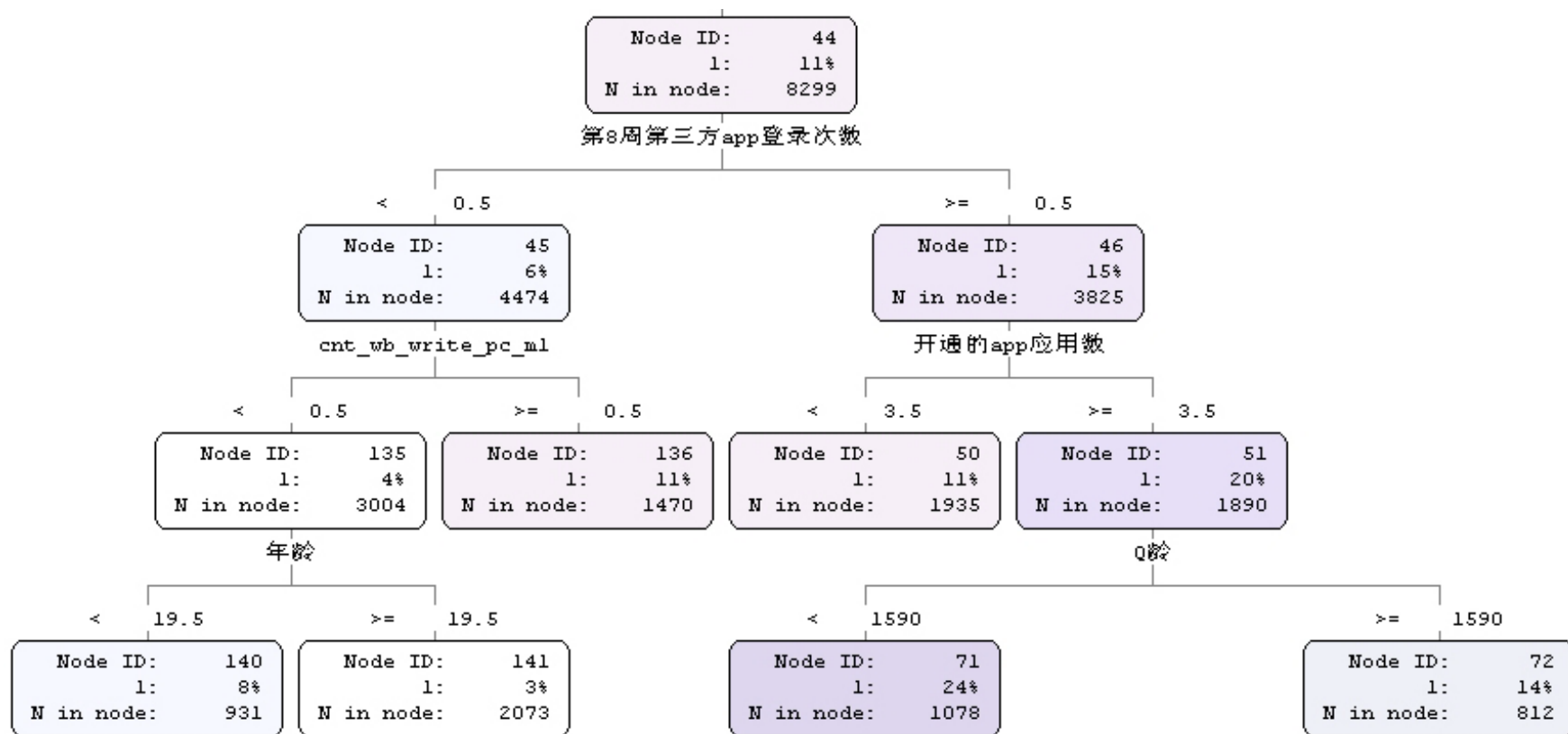


# 生命周期管理怎么做

---



# 决策树模型



APP only

app+basic

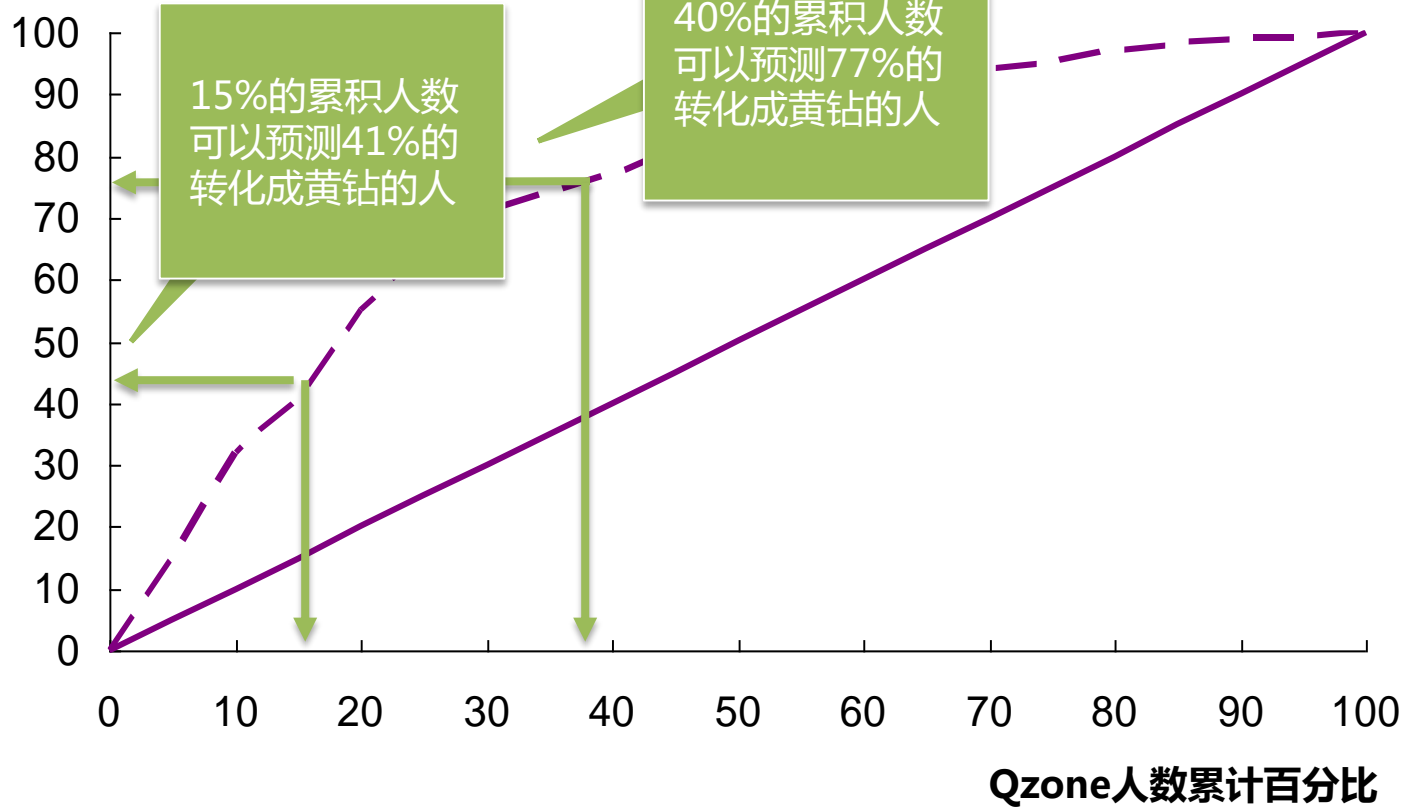
Basic Only

# 预测模型的作用








15%的累积人数可以预测41%的转化成目标用户的人

未来黄钻转化率和Qzone人数累计百分比图  
百分数

未来黄钻  
转化率



# 用户群特征分析

	装扮爱好者	钻族收集群	APP付费	作家	相册狂热者	QQ整体活跃	沉默用户
心理需求							
	喜欢漂亮的东西尤其是装扮类的例如花藤兴趣广泛，比较博爱： <b>活跃于多种活动</b> ；喜欢购买新颖和有吸引力的东西。 <b>年龄比较年轻</b> ，朋友数量和大多数用户相当	拥有 <b>3个钻以上</b> ；喜欢在第三方游戏或者应用上花钱，但对活动有特别偏好：仅在装扮和音乐盒等少数产品上活跃。该人群 <b>男性比较多</b>	他们 <b>喜欢互动娱乐，愿意为此花钱</b> ；最近活跃次数是越来越高的，但是不大使用Qzone基础功能，对他们而言，有趣的游戏是关键。他们往往比一般用户年龄大，朋友较少，男性居多	喜欢在 <b>说说日志</b> 上表达观点，也喜欢在自己和朋友页面上分享、发表，和在别人的页面上评论； <b>时刻都要跟朋友连在一起</b> ，手机功能活跃；不太关心相册。这个人群有很多朋友	非常 <b>喜爱相册</b> ，尤其是上传照片；对游戏/应用不感兴趣，对装扮也不感兴趣。他们有时候自己写写日志，他们喜欢作为客户看别人的Qzone。这群人一般Q龄较长，朋友较多，相对年龄较大	<b>非常活跃的用户</b> ，而且变的越来越活跃。几乎活跃于所有基础功能，但不愿在单买APP；也比较喜欢买红钻，对绿钻，蓝钻和黑钻等不感兴趣。这个客户群 <b>比较年轻，有很多朋友</b>	<b>总体活跃度低</b> 。基本功能上8周呈下行趋势；Qzone里还有余额。这部分人年龄较大。
细分规模百分比	15%	2%	27%	18%	7%	4%	26%
转化率是平均数的倍数 (转化率)	4X 0.4%	7X 0.7%	3X 0.3%	2X 0.2%	3X 0.3%	8X 0.8%	1X 0.1%
价值百分比	22%	5%	28%	15%	8%	12%	10%



# 深入研究博爱的装扮爱好者发现了他们的行为和需求



## 博爱的装扮爱好者

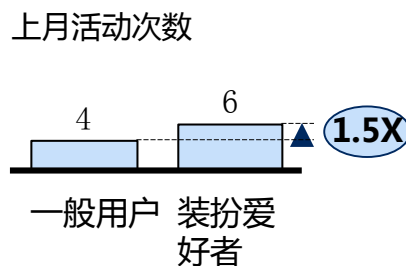
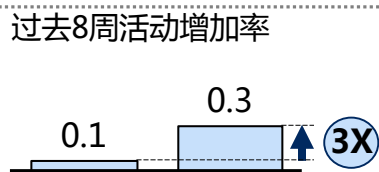
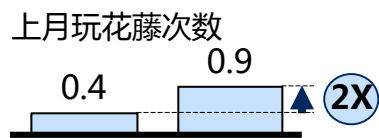
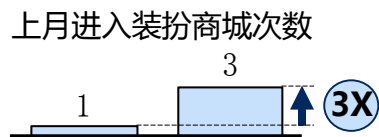
喜欢漂亮的东西尤其是装扮类的例如花藤；兴趣广泛，比较博爱；活跃于多种活动；喜欢购买新颖和有吸引力的东西。年龄比较年轻，好友数量和大多数用户相当

占比大小，%（第一梯队15%的）	15%
转化率	4X
价值，%	22%
男性所占比率，%	55%（vs 平均67%）
平均年龄	22（vs 23）
好友数量	167（vs 168）

## 关键行为

- 登录和保存装扮次数远多于一般用户
- 喜欢花藤
- 他们在装扮上的活动不断增加
- 这些用户兴趣多样，玩Qzone上几乎所有功能，包括Qzone外普及率低的活动和产品（如朋友和微博）

## 装扮爱好者与一般用户比较



## 关键需求

- 希望别人关注，喜欢新事物，如新应用/游戏
- 在Qzone功能上较有经验
- 喜欢看自己和他人的网页
- 有些希望通过Qzone打造理想形象，喜欢参观他人的装扮并与自己的做比较；他们特别不希望在看过的Qzone上留下足迹
- 其他人在基本功能上很活跃，如日志和相册，喜欢装扮。他们的动力在于他们希望别人听到自己、吸引人们的注意力，交更多朋友

# 云标签+生命周期管理

高潜回流濒临沉默

付费高频付费用户

沉默

女性

广告响应

夜猫子

电商用户

高频付费

夜猫子电商用户

女性

广告

未成年

成年用户  
电商用户

付费人数：	30%	↑
付费渗透率：	40%	↑
Arup值：	5%	↑
活动效应：	156%	↑

建模



用户标签



验证服务



功能开发

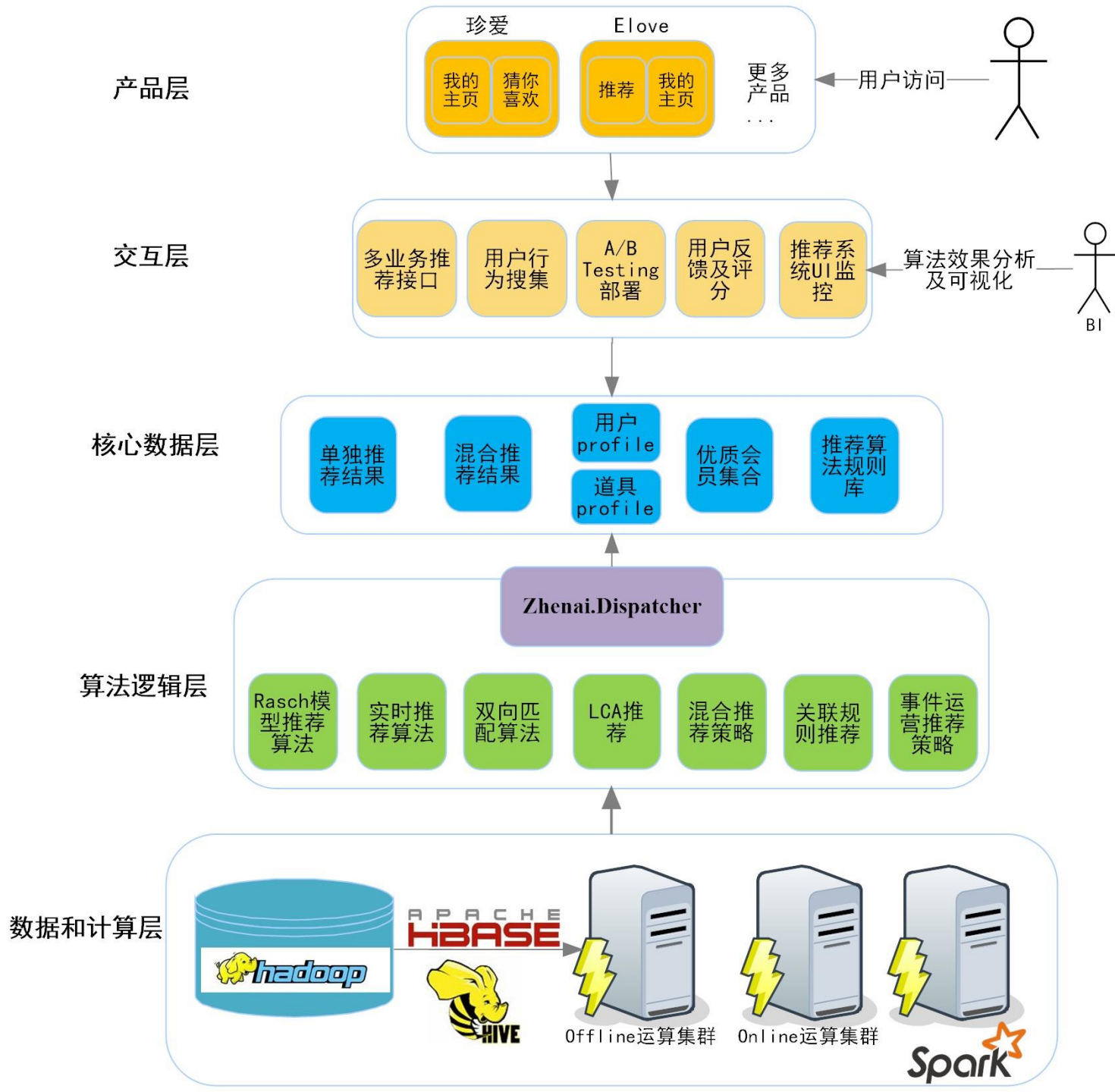


效果跟踪



模型迭代优化

用大数据、数据挖掘能力帮应用提升收入



## 基于大数据机器学习和挖掘的基础设施

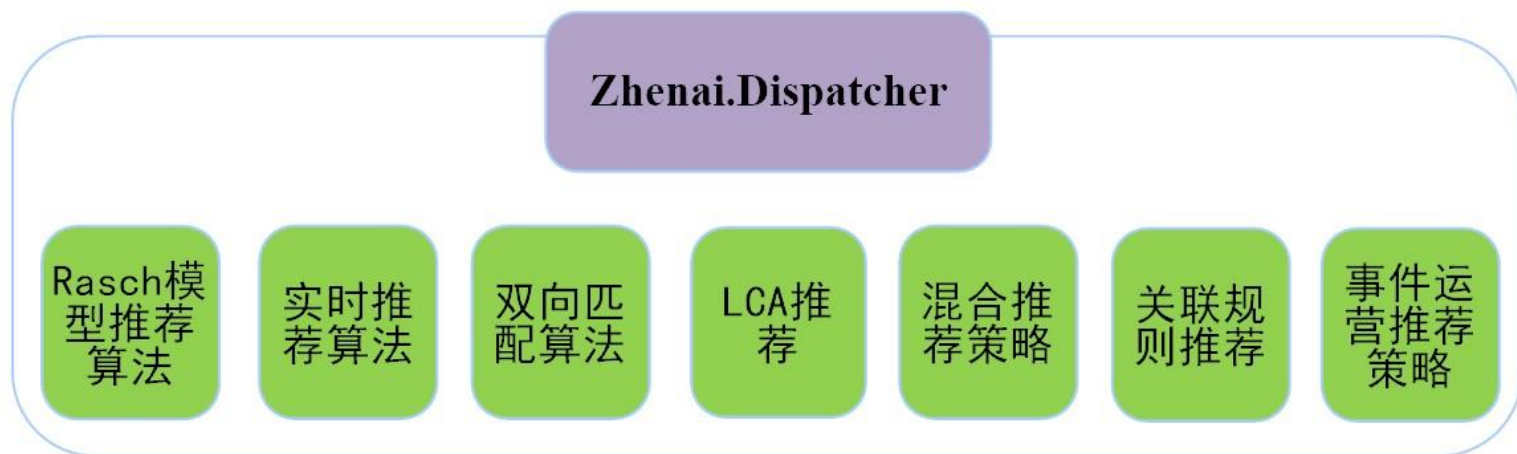


Hadoop分布式运算框架：利用map-reduce统计近PB的用户数据，深入挖掘用户多维度的属性，获得精准的用户偏好。

底层数据存储：利用hive，pig进行数据统计并发布到核心数据层，用hbase存储用户profile的信息，实现更好的可编程特性，满足实时推荐的需求。

Spark计算：满足某些算法对大内存集群运算的需求，同时借助mllib库的强大完整的算法库，大大缩短内存运算的总体计算时间。

## 适合婚恋市场的算法和推荐理念



**算法调度系统：**珍爱调度系统对已有的各类算法推荐结果进行调度使用，根据用户行为和既定分配策略，解决在线推荐算法计算中需要满足SLA对可用性和响应时间的要求。并使用增量学习算法，学习用户偏好并动态调整用户推荐结果。

**双向匹配算法理念：**打破传统推荐系统的“二八”规律，双向匹配以“男女对”为单位计算男女共同得分，同时通过避免某些热门候选人过多的推荐，男女双方产生的推荐都必须在对方的推荐池中并明确不超过一定的推荐数量，解决集中推荐少部分热门用户的难题。

## 男女匹配的神奇方程式

- 1) 基于哈佛等专业心理机构的性格测试和数据模型，来了解用户个性的多个维度；
- 2) 运用测试结果和数据挖掘方法，形成科学的婚恋匹配模式；
- 3) 致力打造“已结婚为目的”的严肃平台，保证用户资料的真实性；
- 4) 打破传统婚恋网站“二八”规律，让每个人都能知道最适合自己的另一半；

男女间性格上的需求匹配  
(如外向, 内向)

男女间硬件条件上的需求匹配  
(如外貌, 资产)

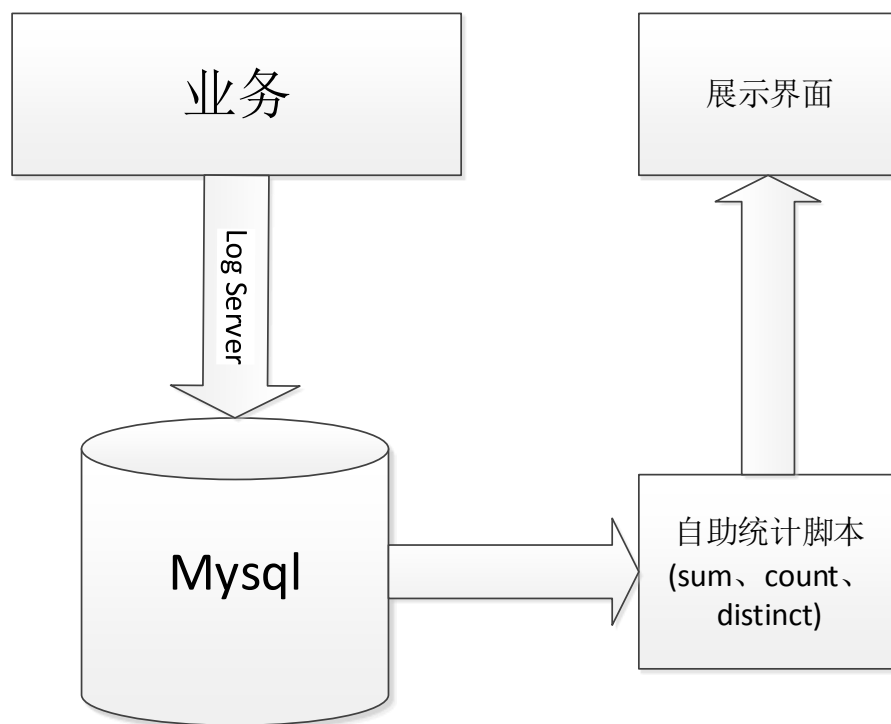
$$r = d_1 |I_i - I_j| + d_2 g_j |k_i - m_j| + d_2 g_i |k_j - m_i| + d_3 |a_i - a_j| + d_4 |b_i - f_j| + d_4 |b_j - f_i|$$

双方对伴侣不可或缺的要求匹配  
(包括物质和心理)

男女间对婚姻的渴望程度

# 数据系统架构——第1代(简单化)

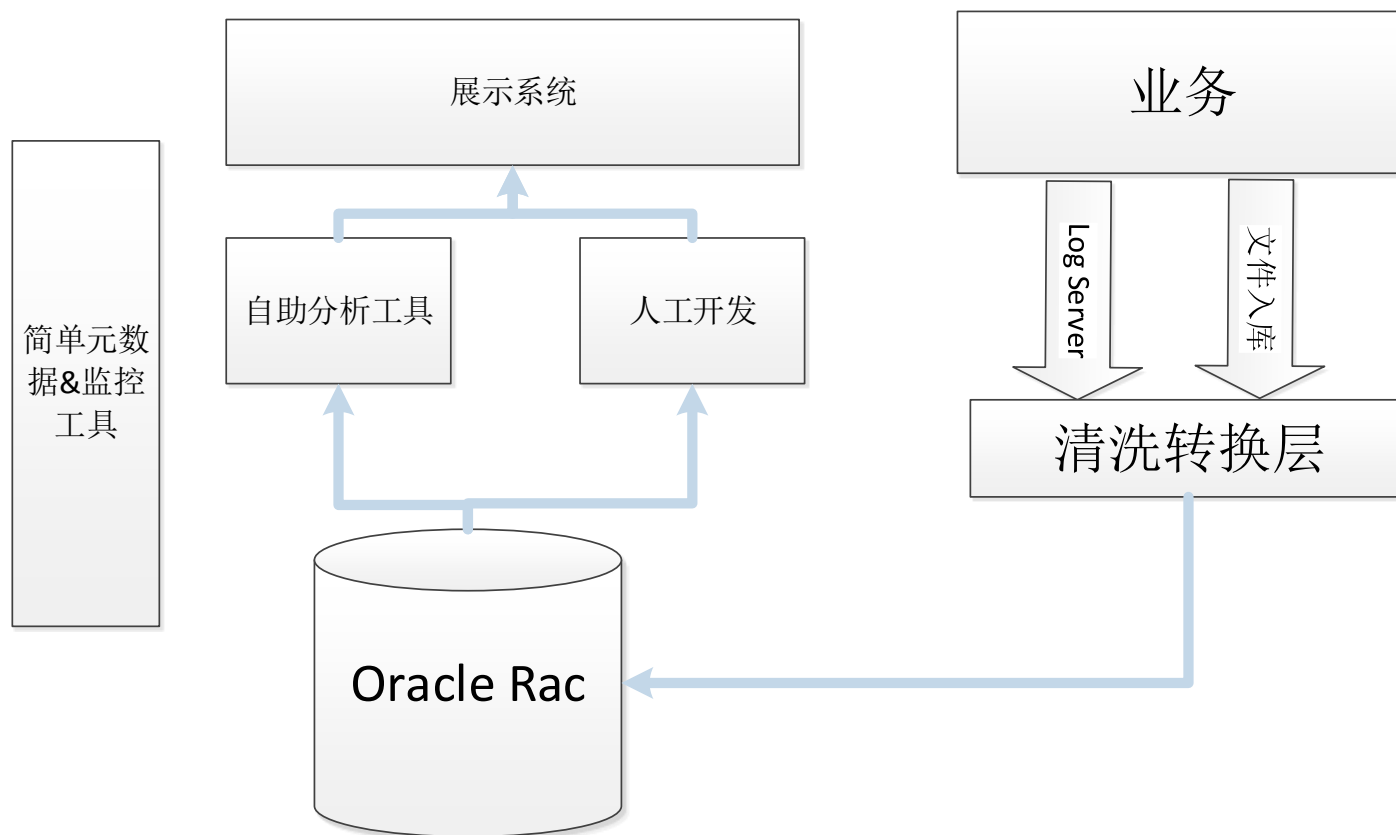
---



- 自助上报
- 业务计数



# 数据系统架构——第2代(集群化)





# 数据系统架构——第3代(工具化)

---

## 展示层工具

- 展示DIY
- 对比工具

## 计算层

- 自助分析
- 指标跟踪
- DIY计算展示

## 存储层

- 自助入库
- 元数据
- 告警监控
- 文件入库

# 数据系统架构——第4代(标准化、开放化)

