Analysys 易观
指数成长的比特动能

如何利用混合云高效做用户画像分析

易观 CTO 郭炜

易观是谁？

Analysys 易观
指数成长的比特动能

易观是中国互联网市场领先的大数据分析公司。自2000年成立以来，易观打造了以海量数字用户数据及专业大数据算法模型为核心的大数据与分析师服务生态体系，并致力于帮助所有拥有互联网产品及服务的企业，洞察自身的产品和用户，对标竞争和市场，并通过对数字用户资产的持续运营，实现增收，节支，提效和避险。

10亿 累计装机覆盖量	66+万 监测的APP数量	1.94亿 移动端月活跃用户	7000+个 数字用户标签个数
8299万 月度新增用户	278个 移动端领域覆盖数	660+个 签约稳定合作伙伴	100亿条 日数据运算量

提纲

Analysys 易观
指数成长的比特动能

1. 用户画像的需求是什么？
2. 混合云做大数据分析的优势
3. 如何利用算法做到精准的用户分布与画像
4. 如何选用优化适合的方案解决用户画像即时查询问题

怎样留住你的用户？

用户从哪里来？流失用户去哪了？怎么重新激活他们？

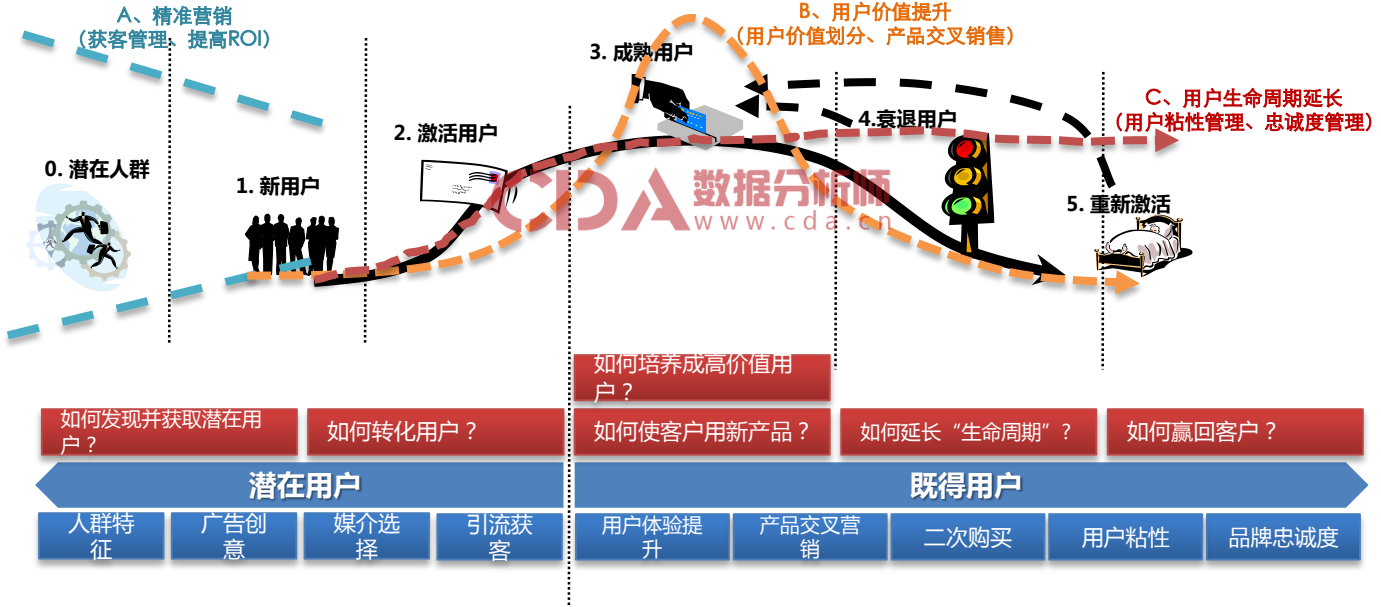
活动一来，用户蜂拥而至？活动一撤，留存效果如何？

流失的客户去哪里了，哪一步流失的？为什么？

怎样激活有价值的沉默用户？如何有的放矢？

用户生命周期数字化管理是精细化运营的基础

Analysys 易观
指数成长的比特动能



用户从哪里来，到哪里去了？

Analysys 易观
指数成长的比特动能

看似简单的漏斗分析，

每个层次都有它的故事：

营销渠道：哪些是你的用户？

浏览运营位：哪些用户看了就走？你的老用户是怎样做的？

提交订单（购物车）：哪些用户冲动消费了？

支付订单：没支付的这些用户到哪里去了？

分享：哪些内容具有传递性？

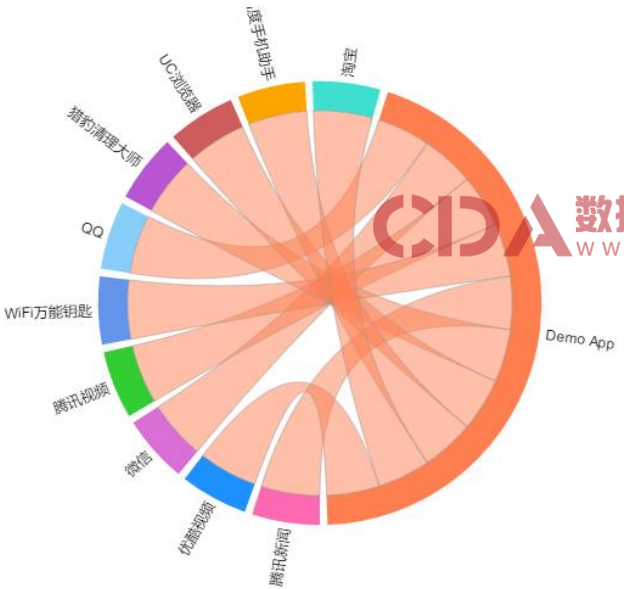
这些用户长什么样？

Analysys 易观
指数成长的比特动能

通过10亿终端的覆盖，通过算法模型计算出你的应用用户特征，性别、年龄...

通过10亿终端的覆盖，通过算法模型计算出你的应用用户特征，性别、年龄...

你的用户从哪里来，到哪里去？



排名	应用名称	关联强度
1	QQ	27
2	微信	13
3	腾讯视频	10
4	WiFi万能...	8
5	腾讯新闻	6
6	猎豹清理...	5
7	UC浏览器	5
8	百度手机...	4
9	淘宝	3
10	优酷视频	3

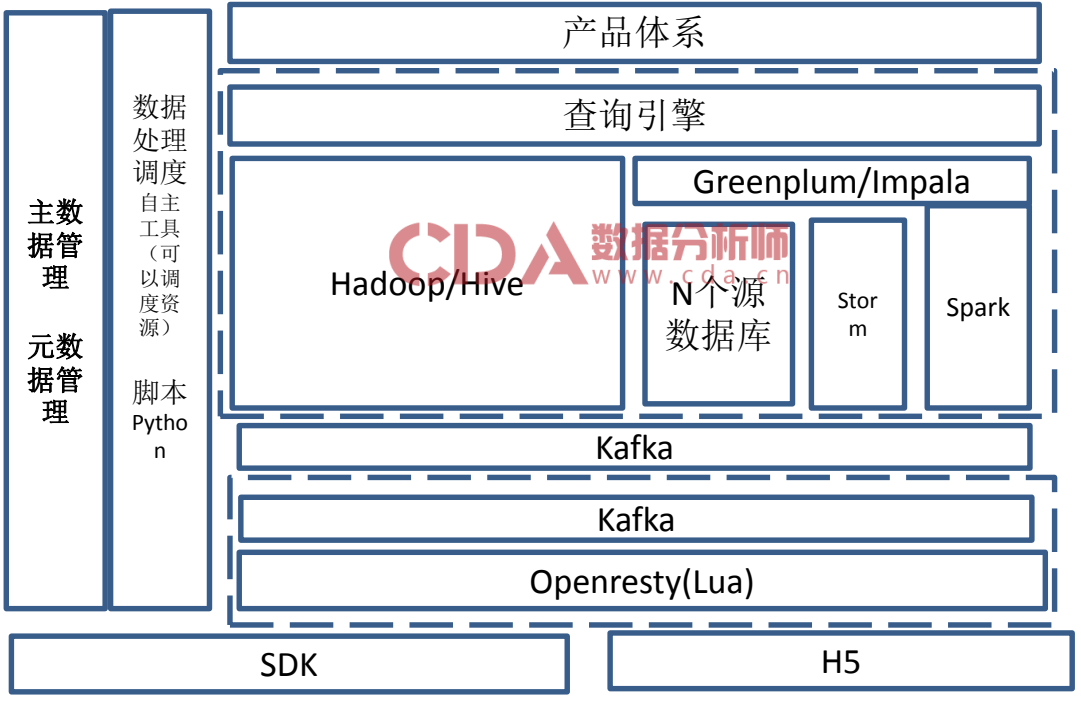
每一个APP都是一个媒体，他们都在哪里？

提纲

Analysys 易观
指数成长的比特动能

1. 用户画像的需求是什么？
2. 混合云做大数据分析的优势
3. 如何利用算法做到精准的用户分布与画像
4. 如何选用优化适合的方案解决用户画像即时查询问题

混合云基本架构



优点

- 底层大数据平台性能稳定
- 云端应用灵活配置
- 攻击防护云端与线下集群可切换
- 灵活的热备极致

缺点

- 需要打通公有云与私有云
- 拓扑结构复杂，问题排查要求较高
- 管理成本比较高，线上线下同时监控

混合云的优势

	大数据云主机	本地+云端混合云
成熟度	国内正在成长	基本成熟
维护	无需维护	需要人管理
性能	磁盘、网络都存在瓶颈	性能高出数倍
稳定性	大批量任务和数据有待提到	稳定
安全	无需防火墙，但会误判	需要防火墙+安全投入
扩展性	易扩展	有购买周期（1个月）

提纲

Analysys 易观
指数成长的比特动能

1. 用户画像的需求是什么？
2. 混合云做大数据分析的优势
3. 如何利用算法做到精准的用户分布与画像
4. 如何选用优化适合的方案解决用户画像即时查询问题

标签如何设置

树形标签不适合用户群使用

正确的标签，应该是图状标签，多种分类多种链接

用户的ＴＧＩ标签

ＴＧＩ：即Target Group Index（目标群体指数），可反映目标群体在特定研究范围(如地理区域、人口统计领域、媒体受众、产品消费者)内的强势或弱势。其计算方法见举例。

ＡＰＰ的标签，映射到用户，通过ＴＧＩ找到用户相关标签

ＡＰＰ如何精准的分年龄

年龄段的划分问题

ＡＰＰ使用时间，ＡＰＰ背后的ＴＧＩ标签序列

分类算法

提纲

Analysys 易观
指数成长的比特动能

1. 用户画像的需求是什么？
2. 混合云做大数据分析的优势
3. 如何利用算法做到精准的用户分布与画像
4. 如何选用优化适合的方案解决用户画像即时查询问题

比较典型的创新企业分析需求

根据不同用户标签，

分析不同APP行为统计

95后，

爱网购的手游狂人，

工作日晚上10:00-12:00，

[illegible]

或者打开网购类APP平均次数？

离散化连续值：

- a) 合并标签（例：95,00,05,10→95后）
- b) 连续值离散化（例>0.85,>0.9）

建立统计对象索引并所有查找范围：

- a) 建立位图利用位的与、或、非操作替代查询条件，找到统计对象集合
- b) 使用ES搜索找到符合条件统计对象集合

使用缩小范围集合找到对象

纵转横，用OLAP或者交并计算提高效率

Analysys 易观
指数成长的比特动能

把每个对象的，90后，95后，00后，购物特征>0.85,购物特征>0.9,手游特征>0.9…标签结构化后存入hbase：结构如右图所示。

比如我们要统计95后，购物特征>0.85且有手游特征的用户群体，我们的筛选条件的向量(0,1,0,1,0,1)构造成一个010101bit数组，和对象的属性做交运算。

例，95后，爱网购的手游狂人，工作日晚上10:00-12:00，最常打开的新闻类APP TOP 10，或者打开网购类APP平均次数？

随机抽样

1. 随机抽样？
2. 如何估算误差，如果？

分层抽样

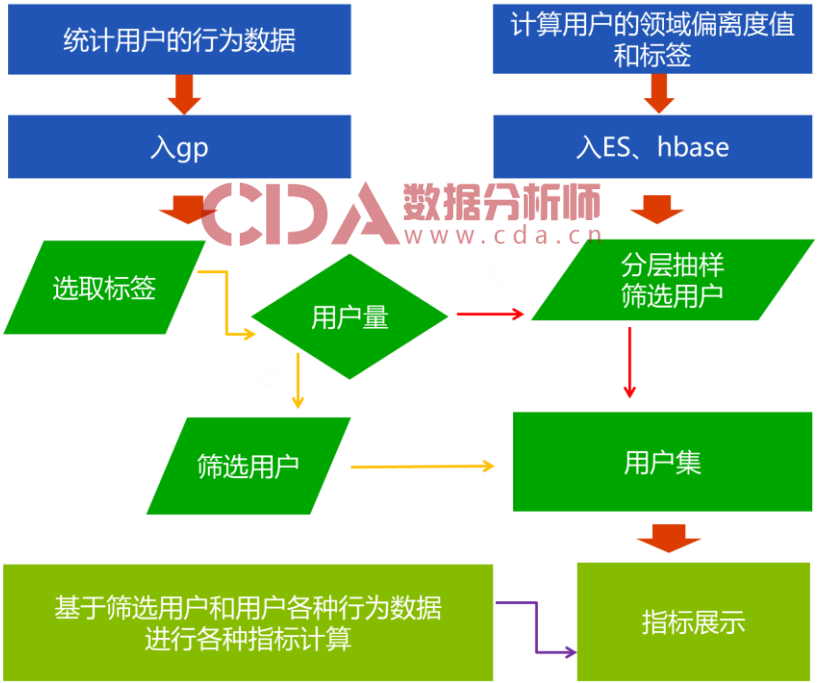
1. 选取整体目标客群
2. 指标领域偏离度，最小客群，以偏离度代替置信，标准偏离度是1，有的人是1.x
3. 以平均偏离度为中心，分别取样，
4. 如何计算最终偏离程度，根据偏离度计算？

领域偏离度通过聚类算法实现分层

用户id	领域偏离度
user_id1	0.5
user_id2	0.71
user_id3	0.9
user_id4	1.3
user_id5	1.5
user_id6	1.76
…	…
user_id2006	3.2
user_id2007	3.3
user_id2008	3.3
user_id2009	3.6
user_id2010	3.6
user_id2011	3.6
…	…
user_id10000000	5
user_id10000001	5.1
user_id10000002	5.1
user_id10000003	5.4
user_id10000004	5.5
user_id10000005	5.5

分层编号	分层值
tgi1	0.8
tgi1	0.8
tgi1	0.8
tgi2	1.5
tgi2	1.5
tgi2	1.5
…	…
tgi10	3.3
tgi10	3.3
tgi10	3.3
tgi11	3.6
tgi11	3.6
tgi11	3.6
…	…
tgi20	5
tgi20	5
tgi20	5
tgi21	5.5
tgi21	5.5
tgi21	5.5

怎样能更快？



分层抽样的误差肯定是比随机抽样的要小。比起全局用户，分层抽样数据量小，计算量小，响应快。

分层抽样不足:有一定的误差，误差平方公式：

$$\mu_K^2 = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2 \frac{\sigma_j^2}{n_j}$$

其中：N表示全体样本数、K表示分组（分层）数、σ_j表示第j组方差、n_j表示j组抽取样本数。

Analysys 易观
指数成长的比特动能

指数成长的比特动能

- 易观千帆
- 易观万像
- 易观方舟
- 易观博阅