

计算广告与流处理技术综述

简介： 案例与解决方案汇总页： 阿里云实时计算产品案例&解决方案汇总 1.计算广告背景 广告仍然是互联网公司的主要变现手段，其市场规模2017年已达3000亿元，据统计全球互联网市值前十的公司广告收入占比高达40%，可见其重要性。

案例与解决方案汇总页：

[阿里云实时计算产品案例&解决方案汇总](#)




1.计算广告背景




广告仍然是互联网公司的主要变现手段，其市场规模2017年已达3000亿元，据统计全球互联网市值前十的公司广告收入占比高达40%，可见其重要性。在这种情况下，与互联网广告相关的技术，我们称之为计算广告，也是最为成熟，市场规模最大的大数据应用领域。




互联网广告领域经过长期发展，分工逐渐精细化，除了各种代理商之外，还出现了ADN、SSP、ADX、DSP等各种平台，市场结构极为复杂，成为了一个巨大的生态。LUMA Partners针对北美市场绘制了一幅全景图，如下。




DISPLAY LUMAscape



  Denotes acquired company  Denotes shuttered company © LUMA Partners LLC 2019

  Denotes acquired company  Denotes shuttered company © LUMA Partners LLC 2019

  Denotes acquired company  Denotes shuttered company © LUMA Partners LLC 2019

  Denotes acquired company  Denotes shuttered company © LUMA Partners LLC 2019

从经济学上看，上面复杂的市场结构可以认为是社会化大生产的产物，而广告的本质其实非常简单，无非是在合适的上下文中寻求受众与广告的匹配，追求媒体、用户和广告主的三方共赢。

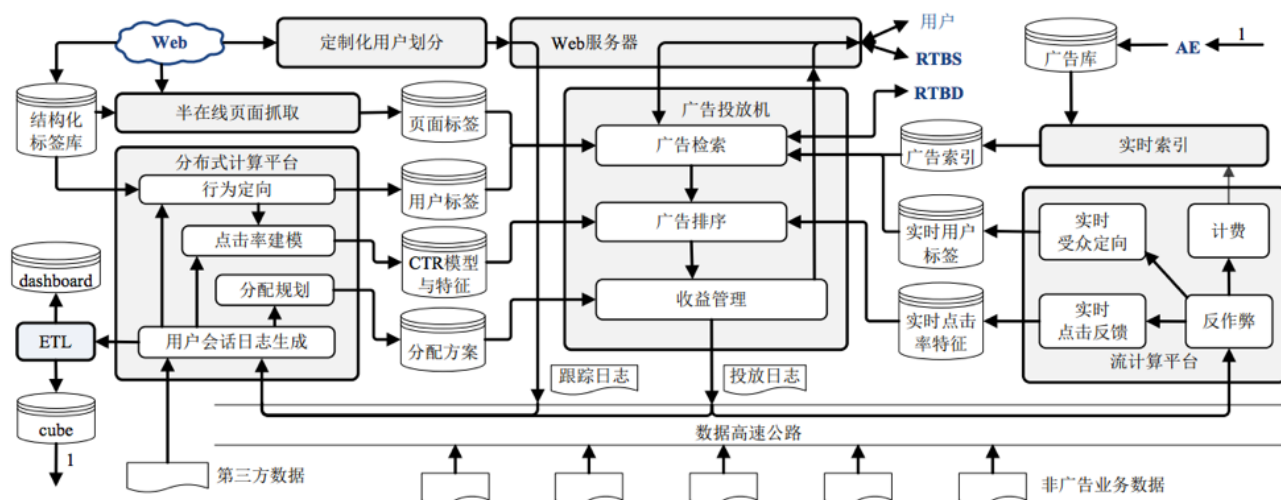


2.计算广告技术架构

互联网广告从诞生那一刻起就与技术紧密相关，这也是互联网广告相比传统线下广告更有优势的地方，通过大数据&机器学习等技术，互联网广告能够在个性化技术的基础上实现更精准的受众定位，不断提高受众与广告的匹配度，让合适的人在合适的上下文中看到合适的广告，达成三方共赢。

而经过长期的发展，计算广告技术已经逐渐成熟，其基本架构图如下：

计算广告系统架构



注：这只是原理性架构图，因为市场结构的复杂性，很多公司可能只有其中若干部分。更为详细的介绍请参考《计算广告》——刘鹏

大概说一下这个架构的流程：

- 用户在媒体浏览网页时，系统会通过一些技术手段抽取当前页面的标签；
- 广告投放机根据用户的标示找到预先做好的用户标签，然后根据用户标签找到对应的适合的广告候选列表；
- 系统根据预先训练好的点击预测模型快速计算各广告的eCPM，eCPM的计算依赖两部分，一部分依赖离线训练好的CTR模型预估点击率，一部分是考虑不同广告的点击价值，综合这两个方面，最终对广告进行排序；
- 最后根据广告分配策略来选出展示的广告，这里不仅仅要考虑本次最优，还要综合考虑整个平台的全局最优；

上面便是计算广告的主流程。

为了保障主流程，系统还需要其他模块。比如离线数据处理模块、在线数据处理模块以及数据通道。

每次投放产生的日志，对接的其他数据以及自己抓取的数据均通过数据通道传给离线数据处理模块和在线数据处理模块。

大家可以看到广告投放的基本流程是确定的，但每次决策会随着场景的不同有不同的结果，其决策依据便来源于离线数据处理模块和在线处理模块，可以说这两部分是整个计算广告系统的神经中枢，也是决定计算广告系统效果的关键部分。

2.1 离线数据处理

前面提到了离线数据处理和在线数据处理都是为了决策者服务，而决策者包含两类，一类是人，一类是机器。这部分的技术主要是离线大数据技术，比如Hadoop、Spark、Hive等。

2.1.1 为人决策提供服务

为人决策服务的，便是大家所熟悉的传统大数据处理技术，如数据仓库等，产出报表、业务大屏或提供OLAP分析服务；

关于这部分可以参考[数据仓库介绍与实时数仓案例](#)

2.2.2 为机器决策提供服务

这部分直接服务于广告投放主流程，从流程中可以看到包括：

- 受众定向，计算受众标签；
- 上下文定向，判断当前环境；
- 点击率预估，相当于寻找受众跟广告的匹配度；
- 分配规划，平衡广告主利益，最大化平台的收益；

这就覆盖了广告过程中各个角色，受众、广告主、环境、匹配度、平台收益等。使用的技术包括离线大数据处理技术与机器学习、数据挖掘等。

2.2 在线数据处理

整个广告的决策是一个在线的过程，传统的离线技术有时候很难满足，这时便单独抽取出了一个在线处理模块作为补充。这部分的技术主要是流处理技术，如实时计算（Flink）。

在线处理环节主要包括：

- 在线反作弊，广告本质上是在卖流量，那么流量作假便能直接获取收益，据ANA统计，大约有37%的在线广告点击是假的，可见在线反作弊模块的重要性，效果不好将造成巨大资金浪费。
- 在线计费，很多广告系统是程序化交易（如DSP），每次点击都会扣除广告主相应的费用，这要求系统能够快速地完成结算，扣除费用，并下线费用不足的广告。计费需要扣除作弊流量。
- 在线受众定向，受众定向主要是计算用户的各种标签，有时用户短期内的行为更有参考价值，产出的短期标签更有效，比如受众突然看到某篇文章进而对某类产品产生了兴趣。这在效果类广告上更加明显。关于长短期兴趣标签可以参考[基于实时计算（Flink）打造一个简单的实时推荐系统](#)
- 在线点击反馈，可以根据用户在线点击情况去调整CTR模型以更好的预估点击率。
- 实时索引，广告是一种商业行为，广告主会根据当前广告的效果调整广告策略，那么每次调整后都需要尽快生效，否则将有可能造成资金浪费，所以需要实时把广告的更新或发布都建到广告索引中去。
- 实时广告链接检测，根据访问日志快速判断某些广告链接是否失效，如果失效则快速将其下线，防止资损。

2.3 离线处理与在线处理总结

其实不管离线处理还是在线处理，本质都是为广告的在线决策和人的决策服务，并没有明显边界。

大数据以离线计算开始，所以很多应用实施在了离线引擎上，但随着在线引擎的发展，其实越来越多的业务都可以在线化，比如离线处理中为人决策的部分，其实可以改造成实时报表，甚至实时数仓，另外在线机器训练也越

来越普及，在线的好处显而易见，可以想象将来在线处理部分会逐步扩张。

3.总结

这里只是概念性的做一个介绍，上面的系统也是一种抽象化的系统，真实的系统可能并不包含上面完整的范围。

比如有些厂商只做纯粹的DSP平台，那么其没有直接用户；还有一些厂商就是单纯的ADN，只有流量，那么其便没有复杂的决策流程；另外还有一些厂商既有ADN又有DSP，或者是既有媒体又有DSP.....等等

但无论如何，从实时计算的角度来看，有几个关键环节是必须要有的，在线流量反作弊、在线计费、在线反馈、在线索引、在线广告链接检测等。

相信随着技术的发展，实时计算版图将进一步扩张。

参考文章：

[数据仓库介绍与实时数仓案例](#)

[基于实时计算（Flink）打造一个简单的实时推荐系统](#)