34 | A/B测试与灰度发布必知必会

2019-01-15 李智慧

讲述：李智慧
时长 08:34 大小 7.86M

在网站和 App 的产品设计中，经常会遇到关于哪种产品设计方案更优的思考和讨论：按钮大一点好还是小一点好；页面复杂一点好还是简单一点好；这种蓝色好还是另一种蓝色好；新的推荐算法是不是真的效果好...这种讨论会出现在运营人员和产品经理之间，也会出现在产品经理和工程师之间，有时候甚至会出现在公司最高层，成为公司生死存亡的战略决策。

在 Facebook 的发展历史上，曾经多次试图对首页进行重大改版，甚至有时候是扎克伯格亲自发起的改版方案，但是最终所有的重大改版方案都被放弃了，多年来 Facebook 基本保持了一贯的首页布局和风格。

相对应的是，一直被认为抄袭 Facebook 的人人网在 Facebook 多次改版举棋不定的时候，毅然进行了重大的首页改版，摆脱了长期被诟病的抄袭指责。但是讽刺的是，事后回头

再看，伴随着人人网改版的是用户的快速流失，并最后导致了人人网的没落，而 Facebook 的守旧却保证了 Facebook 的持续发展。

让 Facebook 放弃改版决定的，正是 Facebook 的 A/B 测试。Facebook 开发出新的首页布局版本后，并没有立即向所有用户发布，而是随机选择了向大约 1% 的用户发布，即这 1% 的用户看到的首页是新版首页，而其他用户看到的还是原来的首页。过一段时间后观察两部分用户的数据指标，看新版本的数据指标是否好于旧版本。

事实上 Facebook 观察到的结果可不乐观，新版本的用户数据指标呈下跌状态。扎克伯格不甘心，要求继续放大新版测试用户的比例，运营团队一度将新版测试用户的比例放大到 16%，但是数据显示新版并不受用户欢迎，数据指标很糟糕。最后扎克伯格决定放弃新版，首页维持原来布局。

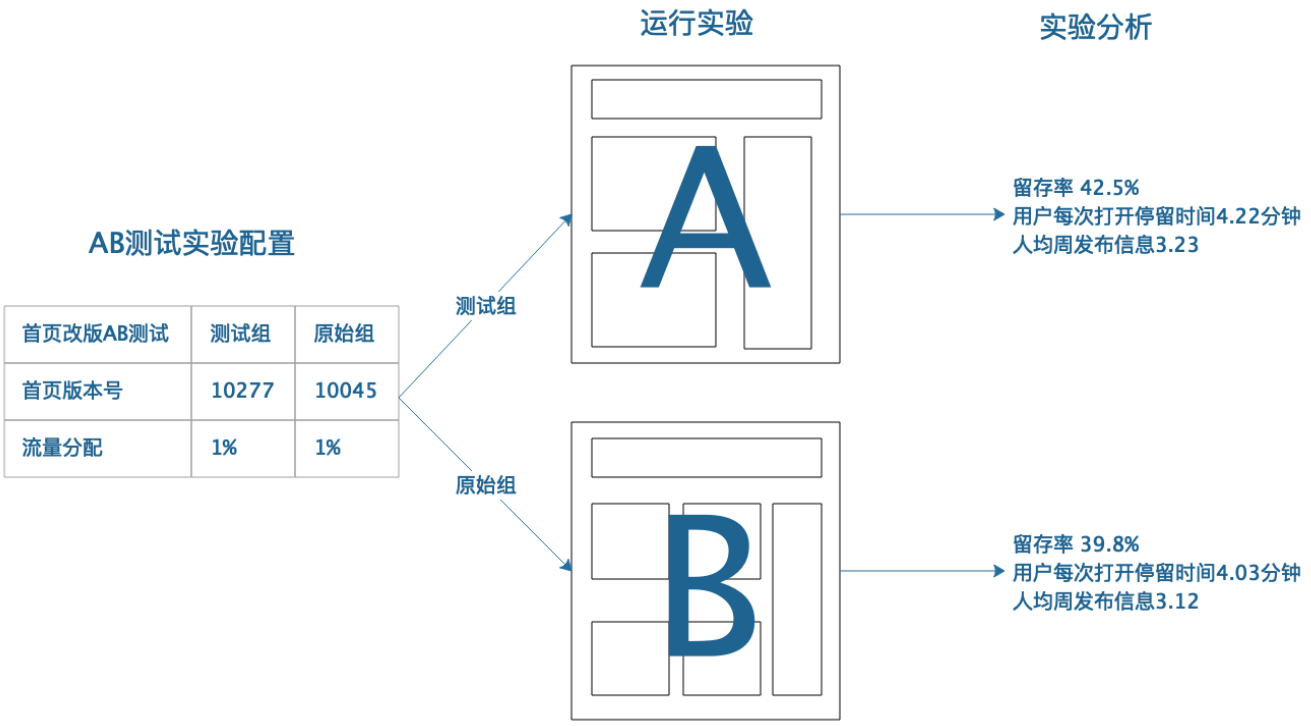
A/B 测试是大型互联网应用的常用手段。如果说设计是主观的，那么数据是客观的，与其争执哪种设计更好、哪种方案更受用户欢迎，不如通过 A/B 测试让数据说话。如果人人网当初认真做 A/B 测试，也许不会贸然改版；据说今日头条为了论证两条新闻之间的分割究竟应该用多宽的距离，同样是做了数百组 A/B 测试。

所以 A/B 测试是更精细化的数据运营手段，通过 A/B 测试实现数据驱动运营，驱动产品设计，是大数据从幕后走到台前的重要一步。

A/B 测试的过程

A/B 测试将每一次测试当作一个实验。通过 A/B 测试系统的配置，将用户随机分成两组（或者多组），每组用户访问不同版本的页面或者执行不同的处理逻辑，即运行实验。通常将原来产品特性当作一组，即原始组；新开发的产品特性当作另一组，即测试组。

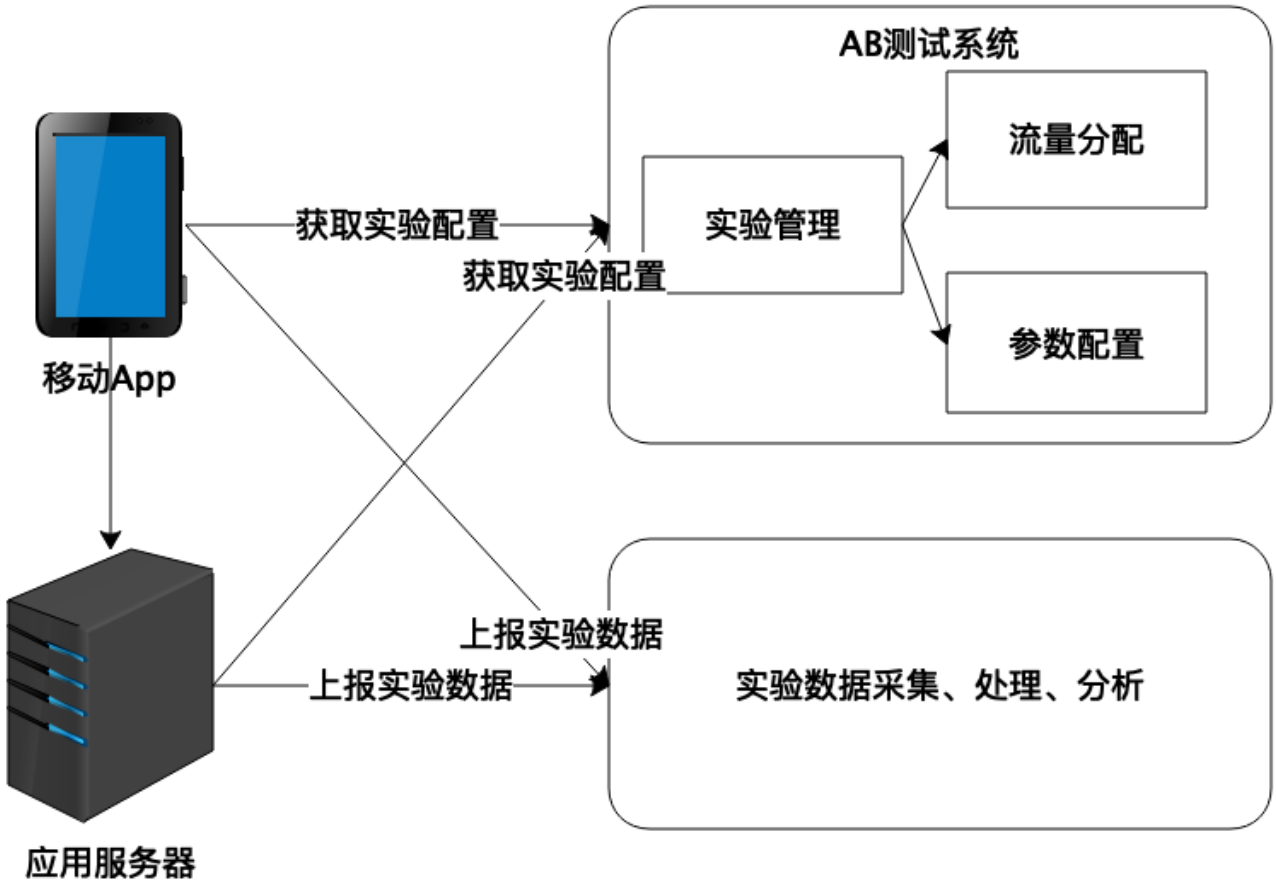
经过一段时间（几天甚至几周）以后，对 A/B 测试实验进行分析，观察两组用户的数据指标，使用新特性的测试组是否好于作为对比的原始组，如果效果比较好，那么这个新开发的特性就会在下次产品发布的时候正式发布出去，供所有用户使用；如果效果不好，这个特性就会被放弃，实验结束。



对于一个大型网站，通常都会开发很多新产品特性，其中很多特性需要进行 A/B 测试，所以在进行流量分配的时候，每个特性只会分配到比较小的一个流量进行测试，比如 1%。但是由于大型网站总用户量比较大，即使是 1% 的用户，实验得到的数据也具有代表性了。Facebook 拥有几十亿用户，如果 A/B 测试的新特性对用户不友好，那么即使只测试 1% 的用户，也有几千万用户受到影响。所以，在进行 A/B 测试时对实验流量和特性的选择也要谨慎对待。

A/B 测试的系统架构

A/B 测试系统最重要的是能够根据用户 ID（或者设备 ID）将实验配置参数分发给应用程序，应用程序根据配置参数决定给用户展示的界面和执行的业务逻辑，如下图。



在实验管理模块里进行用户分组，比如测试组、原始组，并指定每个分组用户占总用户的百分比；流量分配模块根据某种 Hash 算法将用户（设备）分配到某个实验组中；一个实验可以有多个参数，每个组有不同的参数值。

移动 App 在启动后，定时和 A/B 测试系统通信，根据自身用户 ID 或者设备 ID 获取自己参与的 A/B 测试实验的配置项，根据配置项执行不同的代码，体验不同的应用特性。应用服务器和 A/B 测试系统在同一个数据中心，获取实验配置的方式可以更灵活。

移动 App 和应用服务器上报实验数据其实就是传统的数据采集，但是在有 A/B 测试的情况下，数据采集上报的时候需要将 A/B 测试实验 ID 和分组 ID 也上报，然后在数据分析的时候，才能够将同一个实验的不同分组数据分别统计，得到 A/B 测试的实验数据报告。

灰度发布

经过 A/B 测试验证过的功能特性，就可以发布到正式的产品版本中，向所有用户开放。但是有时候在 A/B 测试中表现不错的特性，正式版本发布后效果却不好。此外，A/B 测试的时候，每个功能都应该是独立（正交）的，正式发布的时候，所有的特性都会在同一个版本中一起发布，这些特性之间可能会有某种冲突，导致发布后的数据不理想。

解决这些问题的手段是灰度发布，即不是一次将新版本发布给全部用户，而是一批一批逐渐发布给用户。在这个过程中，监控产品的各项数据指标，看是否符合预期，如果数据表现不理想，就停止灰度发布，甚至进行灰度回滚，让所有用户都恢复到以前的版本，进一步观察分析数据指标。

灰度发布系统可以用 A/B 测试系统来承担，创建一个名叫灰度发布的实验即可，这个实验包含这次要发布的所有特性的参数，然后逐步增加测试组的用户数量，直到占比达到总用户量的 100%，即为灰度发布完成。

灰度发布的过程也叫作灰度放量，灰度放量是一种谨慎的产品运营手段。对于 Android 移动 App 产品而言，因为国内存在很多个应用下载市场，所以即使没有 A/B 测试系统，也可以利用应用市场实现灰度发布。即在发布产品新版本的时候，不是一次在所有应用市场同时发布，而是有选择地逐个市场发布。每发布一批市场，观察几天数据指标，如果没有问题，继续发布下一批市场。

小结

A/B 测试的目的依然是为了数据分析，因此通常被当作大数据平台的一个部分，由大数据平台团队主导，联合业务开发团队和大数据分析团队合作开发 A/B 测试系统。A/B 测试系统囊括了前端业务埋点、后端数据采集与存储、大数据计算与分析、后台运营管理、运维发布管理等一个互联网企业几乎全部的技术业务体系，因此开发 A/B 测试系统有一定难度。但是一个良好运行的 A/B 测试系统对企业的价值也是极大的，甚至可以支撑起整个公司的运营管理，我们下期会详细讨论。

思考题

A/B 测试需要在前端 App 根据实验分组展示不同界面、运行不同业务逻辑，你有没有比较好的设计方案或者技术架构，可以更灵活、对应用更少侵入地实现这一功能？

欢迎你点击“请朋友读”，把今天的文章分享给好友。也欢迎你写下自己的思考或疑问，与我和其他同学一起讨论。

上一篇　33 | 一个电商网站订单下降的数据分析案例

下一篇　35 | 如何利用大数据成为“增长黑客”？

精选留言 (9)

写留言

菲戈
2019-01-16

6

为什么讲大数据的课程，会说到A/B测试去

展开

作者回复: 原文：A/B测试的目的依然是为了数据分析，因此通常被当作大数据平台的一部分。

A/B测试是大数据分析和大数据平台的重要组成部分，但是关于A/B测试系统架构的资料非常少，如果说Hadoop、Spark的资料你可以从网上随便搜，那么完整的A/B测试系统架构的资料可能只有这个专栏才有了。
多说两句，Google发表大数据论文距今15年了，Hadoop开源也十几年了，Spark出现也快10年了，如果我们今天学大数据还是眼里只有Hadoop、Spark，真的太OUT了。大数据生态体系包括

Hadoop这样的大数据产品，还包括大数据平台、大数据分析、大数据机器学习，我的专栏是一个关于大数据技术体系的完整知识框架，希望能对你学习大数据起到作用。

Zach_
2019-01-15

3

除了AB实验，还可以提出AA实验，ABC实验的概念

AA实验可以理解成：实验的配置相同，但划分到不同的用户群体

ABC实验可以理解成: 一个实验的多组不同配置而非两组不同配置分别下发到不同群体...
展开

作者回复: 是的，可以根据需求设计实验

null
2019-01-16

2

请问老师，如果AB测试，涉及到调整了数据结构，或者业务逻辑较大改动，是否还有用呢？比如统计中需要全量数据，AB测试分成两个不同表来存。暂时考虑的是冗余存储比调整报表逻辑好，但是不知道是否会影响到AB测试的结果，毕竟有一部分是多做了近一倍的事，性能、用户感受这些指标结果可能又不准确。
展开

作者回复: A/B测试可以理解成在原来的打点基础上增加了实验ID和分组ID，数据存储和结构跟原来一样，SQL统计的时候根据ID分别统计，就得到各个实验分组的PV转化率这些指标。

Zach_
2019-01-15

1

看带着过了一遍，我现在觉得AB实验还是很有意思的。

用户请求AB实验成功后，AB后台会下发一组配置给该用户，用户的App会将这组配置作为参数加载进来，
并在下一次请求前，不会改变APP的界面和效果，直到下一次这些AB实验的参数发生改...
展开

程序员小灰
2019-01-15

1

AB测试的逻辑偏复杂、需求也是花样百出，对于SDK，每做一个功能，逻辑设计就要将近一周，代码开发两天。像flurry友盟等单纯数据收集的SDK，很长时间都不会发版。

请问老师，怎么把AB测试的SDK内部逻辑做的比较灵活，目的是适用业务需求变化，还不用频繁发版。

展开

强哥
2019-01-15

1

AB test总体分为三大部分，实验方法，指标计算，效果评估，整体流程还要结合公司的业务，例如流量的划分，指标体系的建设等。APP端一般都是通过sdk进行埋点数据。然后进行etl。

展开

hallo128
2019-01-29

AB测试的核心原理是很简单的，就是统计学中2个总体的比较问题。
难度在于整个系统的自动化搭建，从如何抽样，如何安排试验，但最后数据的传递返回处理。最后才对已有数据进行统计检验。
不过从这个系统涉及到的统计知识会有：试验设计（是否为正交在此阶段考虑），调查抽样，假设检验。...

展开

小老鼠
2019-01-22

AB测试用户喜不喜欢是如何获得的?

展开

作者回复: pv uv 留存各种数据指标下降了，就是不喜欢

hxppk
2019-01-18

abtest 流量分配环节，如何做到百分比流量分桶，同时也做到用某些event条件等划分流量，让流量高效利用？两种划分逻辑如何共存？

作者回复: 流量划分需要尽量随机，保证实验结果客观，不应该有太多的划分方式。