数据分析师必学四大精髓

GrowingIO ·

2016-09-01

定目标、用产品、看数据、记指标，你也可以成为牛逼的数据分析师！

本文作者：单元明，GrowingIO联合创始人、产品VP。单元明毕业于复旦大学和华盛顿大学，先后就职于Coursera、LinkedIn和Rocket Fuel，主要从事互联网产品和移动分析、用户增长和货币化等方面的工作，有多年数据化业务驱动经验。原文发于GrowingIO技术博客，授权36氪发布。

作为一个数据分析人员，有没有经常被业务人员抱怨报表出的太慢、被工程师嫌弃埋点沟通不精准、甚至被老板怀疑并没有创造什么商业价值.....

好好学习这四步分析精髓，从树懒慢先生变成一部行走的AlphaGo，真正的人工加智能，数据分析又快又准！让你一举成为公司头牌，用数据驱动业务增长，快到飞起来！

直接上干货！

一、定目标

首先要设定好业务目标，不但能明确接下来分析究竟是为了什么，而且在人人都加班加点不怕猝死的环境下，可以尽可能的优化时间，放到最能产生价值的地方。

比如下面这些:

1. 每次开会前有明确的目的
2. 接受任务时想过真正的需求是什么
3. 定义指标时想过以后怎么用
4. A/B测试前都设好成功标准
5. 制作报表前想过结论是什么，推荐什么行动
6. 分析产品时是要让用户更满意，更多付钱，更长时间使用，还是更经常使用

同时，作为数据分析师还要注意团队合作中可以实现高效沟通，让搭档和老板清楚了解回报是什么，为什么要支持这个决定，为什么要分配资源，为什么要放下手中别的活，清晰的目标就是沟通的核心，价值就是让对方眼睛发亮的关键，至于背后处理了多少数据，提交了多少SQL，估计了多少模型参数，制作了多少漂亮图表，没人在乎。

数据分析，结果导向。做事前先想想要创造啥价值，最好能直接产生业务增长，满足人的

欲望。**没有目标，光在那炫酷玩屠龙术，迟早被老板砍掉，被客户踢开。**

二、用产品

培养分析师的用户视野

定好了分析目标，就要从数据分析的基础，用产品起步了。**用产品可以帮助分析师获取缺失的信息，培养同理心，站在用户的角度看问题。**

数据的颗粒度再细也是有限度的，而且是冷冰冰的，反应的是抽象后的状态。当看到一个注册漏斗有很大比例用户离开时，说明了什么？数据上也许说明了有需要改进的地方，但注册流程中究竟哪里需要改进？用户到底想要什么样的体验？用户现在的感受如何？这些形象具体的信息，或者说没能记录的“数据”，就需要按照用户的方式，实践几遍，才会有所感悟。

当自己试着去注册，却发现输入几次密码死活通不过，是不是自然而然心中跑过一万头草泥马？为什么至少要8个字符？为什么又要大写，又要小写，还要数字，特殊字符？又不是银行帐号。而且密码要求怎么不明确显示出来？

更要命的哪些算特殊字符可能也没不说清楚，用户估计槽都不想吐了。亲自使用产品，通过眼睛，耳朵，大脑，和心，体验那些被数字抽象掉而流失的感觉，也就明白了用户的沮丧，改哪里，怎么改也就变得清晰。而这一过程，也培养了分析师多角度观察问题的能力，成为连接用户和公司的桥梁。

理解数据定义和业务逻辑

数据如何生成？传统方法是靠工程师写事件处理函数。数据需求哪里来？业务人员要看得指标。那么触发条件是什么呢？业务和技术的理解是一样的么？销售说，我想知道这个广告位点击率多少。工程师说，广告位的点击次数和浏览数已经有了。马上就拿这两个数据除一下就去给销售么？销售是怎么定义点击率的？分子分母分别是什么含义？浏览次数是指眼睛看了多少次么？如果亲眼去看下广告，可能发觉广告位只要渲染了，即使未在显示器中出现也算。这一进一出，点击率翻个倍都有可能，工程师知不知道销售的确切意思？销售知不知道工程师的实现？如果客户问起如何回答？

通过产品使用，才能体会到数据整个上下游中的定义问题，增强一致性，而不是默认工程师给的就是业务需要的。随着无埋点技术的发展，现在数据分析已经能做到由业务人员直接进行数据定义，而无需工程师过多参与。

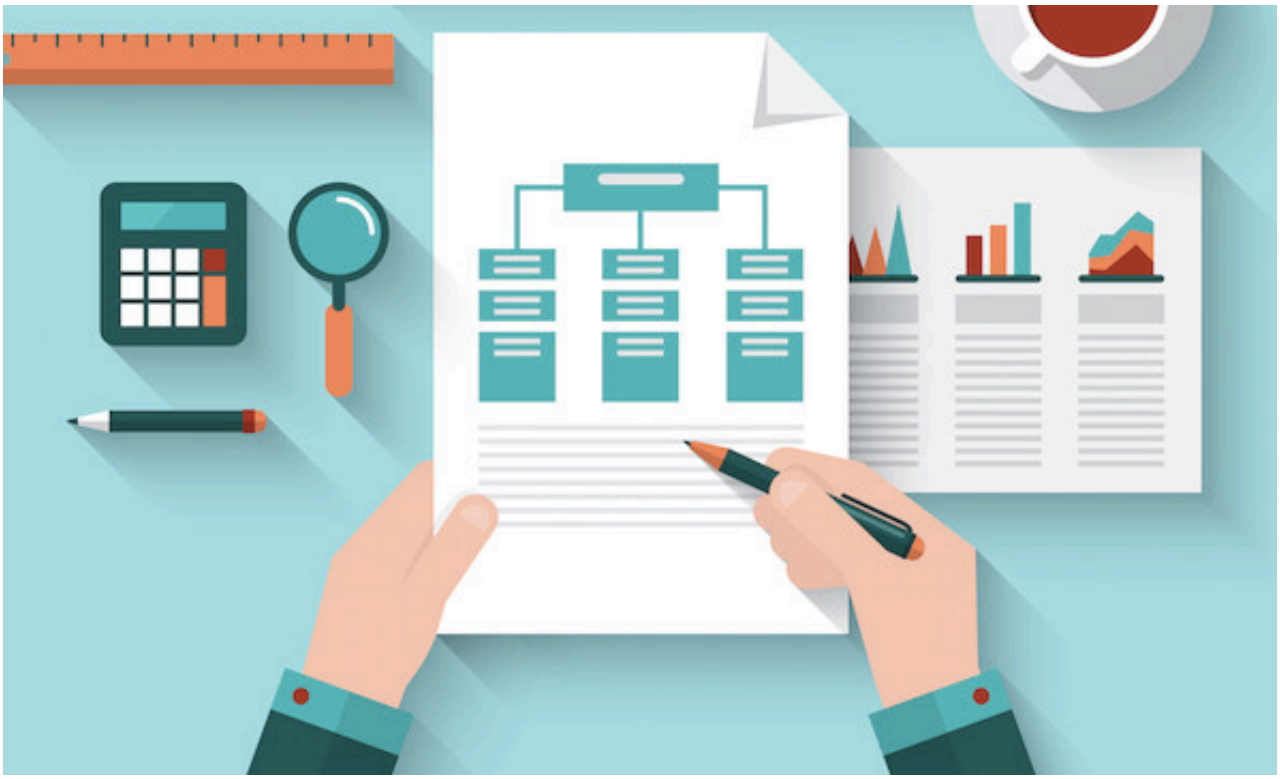
但在定义中，仍会涉及到具体的定义细节，只有通过实实在在的使用产品，才能体会这些细节上的差异和因此带来的蝴蝶效应。

体会细节差异

除此之外，用产品还可以快速检查数据质量和边界条件，特别适合高速迭代。点下按钮会否生成重复记录？有没有漏掉的触发？输入框最小几个字符，最大有没有设限，哪些情况会产生错误数据，定义的标识符是否正确记录，用下产品，看看下游接收到的数据，马上就能知道。

三、看原始数据

有明确的目标，也了解自己的产品，还需要什么？统计？编程？算法？忽悠？这些当然重要，但更重要，也往往更容易被忽略的是建立起对数据本身的感觉。**对数据的感觉，说穿了，就是对数据结构，具体数据形式的熟悉和敏感程度。**



数据分析师需要通过看原始数据找感觉

列举些常用的方面如下。

表格

常用表的作用；

数据库总体数据量；

每天会增加大约多少数据量；

一个时间范围内的全表扫描需要的大致时间；

是否需要采样，或安排到半夜进行事务数据和分析数据的误差；

有否超过允许范围合计表的数据源；

合计逻辑是什么更新频率和相对现实交互的延迟映射表的关系，一对一，一对多，还是多对多记录的是快照，还是所有历史变化可用索引原始数据的结构；

哪些信息是键值形式，哪些是数组形式原始数据留存政策和时间。

列

常用列的意义；

列之间的关系列值的分布；

不对称情况，是否合适作划区哪些有效，哪些过期的；

哪些有问题的如果是枚举值，常用的值，代表什么意思是否有0，负值，空值，特殊值的排除和处理时间存储的形式；

UTC还是本地时间，单位是秒，毫秒还是微秒级字符串中可能含有的列分隔符；

乱码值是否应该独特唯一、是否做到独特唯一数据类型，显示的都是数字，但是否错误的存成字符串

其他

产品迭代中，新旧数据的区分点；

不同的业务逻辑线下上传；

第三方的离线数据源；

常用的黑名单、白名单、测试名单；

授人以鱼不如授人以渔，看原始数据，就是建立对数据感性认识的最好方法。

让分析师沉下去，了解公司数据生态圈的各种主要细节，从而能高效产生新的聚合信息。而不是浮在上面，只知道一些归并抽象后的现存量化值。需求千变万化，总有很多情况，没有可以直接使用，合计好的表格，这时就需要去建新的业务逻辑，生成新的合计表，对数据细节的高度把握，对流畅完成这一过程有很大帮助。

另一方面，挖掘洞察的过程中，很大一部分时间都是在搞数据清理：检查正确性，去除污

染，转变成可用形式等等。

而对数据的熟悉程度就直接影响这些工作的效率。而每当需要记录新的跟踪，也能知道新信息加在哪里更利于使用。

日常工作中，推荐两个方法去熟悉原始数据，一是根据实际需求，去观察相应的数据来培养感觉。二是可以有意识的，刻意的投入一些时间去看当前职责之外，但和公司主要业务产品紧密相关的各种表格，各种原始数据的具体内容和形态，会对以后的工作产生很大帮助。

四、记业务指标

作分析除了需要各种软硬通用的能力，比如统计，编程，算法，忽悠等，还得对公司的业务非常熟需。从分析的角度看，增加业务的熟悉程度最直接有效的办法就是记指标。如果不看报表，下面这些问题有几个能立即回答出来。

数据分析师的基本功：熟记指标

常见的业务指标

公司平均每天 / 每周 / 每月营业额，活跃量，流量大小；

周末和周中关键差别和特征；

早上，中午，晚上用户关键差别，活跃数，流量；

北京和上海，各主要地域的市场份额，消费能力，平均每用户营业额；

公司下个季度预期增长率，预期今年的营业额；

桌面和移动的活跃比例，收入比web和app的比例免费用户和付费用户比例、主要差异、80%的营业额由前百分之几的用户提供；

主要漏斗，如注册，登录，付费，提交等，每一步的转化率，流失率主要产品的客户留存；

获取用户的成本，用户的生命周期价值公司平均每天 / 每周 / 每月营业额，活跃量，流量大小周末和周中一般差别；

以上这些指标，很多互联网公司都有，但能不能记住，是区别一个分析师水平的重要方面。好的分析师，这些都烂熟于心，几乎成为了第二本能。

熟记指标的优势

熟记关键的指标，在看到异常波动时，才会敏感的察觉有地方不对，也就是我们通常说的“感觉”。这一点在公司人与人交互中尤其有效，因为交互是实时性的，需要有立即的反应。

比如开会中讨论新的方案，需要立即指出可能存在的问题，并给出质疑的原因和证据，引导会议成员的思路并提出解决方案。而如果对业务指标不熟悉，很难有这种感觉，或者就算有所察觉，但因为不够熟悉不够自信，就需要去翻看报表，找到相应的关键指标，前前后后可能需要十几分钟。这在一个人做分析时，没有太大问题，但在会议，讨论等实时性很强的互动中，显然是不合适的。

熟记指标另一个巨大优势是能给分析师带来巨大的可信度。

分析师相比其他职位最大的优势是能接触到数据，海量的数据。在规划战略，定位产品时，很多观点都是基于逻辑推理，行业经验，类比假定，而分析师就有机会提供更加量化的指标，为合理的观点提供强有力的支撑。所谓事实胜于雄辩，“我们随机抽样，90%用户支持现在的定价”就要比“一般大家都这个价位”要有说服力的多。长期进行以量化事实为依据的交互，分析师能赢得很多的信任，从而更有效的领导跨组合作。

如何记住业务指标

别小看指标的数量，虽然是大数据综合提炼而成的统计表征，但指标自身可能也是“大数据”。拿活跃用户这一个指标举例，看过去7天按每天整合，就有7个数据点。如果再按地点北京上海等分类，可能又有10个点，然后再加上设备，渠道，付费级别，参与程度，访问来源等维度，以及互相之前的环比，同比等等，那一个指标变成几千或者几万个数据点轻而易举。如果要把这些全记下来，那基本不用干别的活了。

比较有效的方法就是抓大放小。

首先，如果日活是1234567人，那么后面那些具体数字基本上没有太大意义，记一个120万人就可以了，要得是那种大致的感觉，不是银行出纳分厘不差的精确。

其次，各种维度记几个主要值就行，比如地区就记北京，上海，广州等。而设备就是安卓，苹果，桌面等，不需要背黑莓，win phone等各种小众移动的份额。

最后，优先记整体情况，只有一个维度时的聚合, 如果有时间再看多维度的交叉细分。比如，N天前，北京、苹果、付费用户，这4个维度交叉后的指标可能也有价值，但把不交叉时的主要时间、地点、平台、用户类别的指标记清楚，覆盖范围要广的多，记的数量也要小得多。如果你能定下清晰地目标、熟练使用产品、熟悉原始数据、熟记关键业务指标，恭喜你已经从一名数据分析人员进阶为一名合格的数据分析师了。

如果你还会能掌控最新、来自硅谷最前沿的数据分析产品，比如无需埋点、全量实时采集的新一代数据分析产品；恭喜你，不只是一名合格的数据分析师，简直就是一名数据科学家啦。