查看: 18663 | 回复: 142

[面试经验] 发一波DS面试准备经验总结回报地里

[复制链接] | 试试Instant~

feiwudefeng 发表于 2019-2-18 07:22:03 | 只看该作者

精華

垄头 电梯直达

本楼： 100% (1) 0% (0)
全局： 100% (217) 0% (0)

2019(1-3月) 分析|数据科学类 硕士 全职@GoogleAirbnb, Thumbtack - 内推 - Onsite | Other | 在职跳槽

楼主从去年10月份开始系统复习准备DS analytics方向的面试，在准备过程中在地里学习了很多前人的面试经验，也被推荐了很多很有效的资源，非常喜欢这个论坛的气氛。现在找工结束，想发一些自己准备面试以及面试的经验（心得和反思都有）来回报地里。因为签了NDA，不太方便透露具体的面试题还请大家见谅，但是会举一些自己在面试中遇到的问题和自己的思考，欢迎大家一起探讨。

先报一下自己最近这波找工作的情况，因为比较忙所以没有海投，不然连猎头电话都打不过来了：

简历拒：uber，pinterest，snapchat，dropbox，glassdoor

店面挂：linkedin（技术二面）

onsite挂：quora，microsoft

onsite加面挂：facebook（对不起我热心内推积极提建议的舍友啊TAT）

offer：thumbtack（PA），google（BA和PA两个ladder），airbnb（DS-Analytics）

和很多小伙伴一样，我这次主要的方向是analytics track，准备的方面也是依据这个track的要求进行的，分为以下几个方面：

（1）Product Sense：

地里被很多人推荐的cracking the PM interview里关于behavior的章节我看了三遍，estimation，product和case三个章节我看了至少五遍吧，最开始的时候几乎每面一个公司的每一轮面试前都会看一遍，听上去看了挺多遍的其实看到后几遍就很快，看到一个标题大致就知道这个部分说的是什么了，主要的目的是过一遍思路，让自己的思维在面试前活起来。我觉得其实很多产品题面试的时候卡壳但是面试后多花时间想一想就有解了，产品题之所以难，在我看来部分原因是思路容易受面试时紧张心态的限制发散不开，面试前过一遍，确定自己脑子里有什么武器至少对我来说还是很有用的。

其他推荐的材料我自己看过的还有case in point，这个因为case太多了我只看了两遍，但其实这本书在解决case study（不是product question）上比cracking 我觉得要简单好用，里面的体系很有说服力，例子又多，不只能让你知道what is working，也能告诉你what is not working。 -baidu 1point3acres

https://learn.1point3acres.com/courses/ds501-analytics/

这个是地里大神小k老师的一个付费课程，我在面试后期才开始用的，但是有点后悔买晚了。里面涉及了很多面试中考察的知识点，

虽然不是每道题都讲得很透，但是如果自己先下苦功夫准备过analytics再来听这个，会非常有拾遗和验证心得的效果。

除此之外，每面试一家公司之前，推荐一定要上地里看面经，一定要在纸上多写几道产品题的答案。我当时面fb的时候在纸上写了25道产品题的自己的答案，面linkedin的时候也准备了15道，虽然最后这两家都g了，但是对我如何把资料和课程里的东西内化起了不可替代的作用，后期面狗家和空气床的时候其实是没有什么面经的，面试的时候也没有面经题，但是前期F和L的准备沉淀下来了，被问到的题自己没见过也没有很虚。

（2）SQL：
这个比较简单粗暴，Leetcode的sql我刷过至少5遍，hankerank我刷过3遍，基本上就没问题了。需要提醒大家的是如果你有一阵子没刷了，一定要再刷一下，手会生的。

（3）Python - manipulation & cleaning：
在这波面试前，我其实是不太会用python做数据处理和可视化的，如果小伙伴中有跟我一样的人，我觉得我用的这些资源真的都非常好，后来面试的时候甚至又不止一次要我现场用电脑处理数据做可视化提建议，用过以下资源表示完全不虚：

（3.1）udemy python for Data Science and machine learning bootcamp：
https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp/learn/v4/t/lecture/5733448?start=0
这个资源我只看了前半部分关于数据处理和可视化的但是看了两遍（第二遍是快进式刷的），非常系统地教了pandas和seaborn

（3.2）datacamp：
里面有关于pandas，matplotlib和seaborn从入门到精通所有相关的课程，我都上了，和udemy相互印证以后，基本上就有了一个比较清楚的概念了。

（4）AB-Testing：
udacity上的那门AB testing的课是入门的利器，第一次上那门课的时候仿佛打开了一扇新世界的大门。但是其实那门课很多细节是没有讲清楚的（至少我自己没听明白），比如没有涉及t test，没有讲清楚variability， bootstrap， A-A test的意义，怎么在实验设计阶段应对learning effect， network effect，和其他一些corner case。 所以我觉得这门课入门很好，但是绝对是不够的。

（5）统计：
我17年初面过一次FB，那个时候自己还挺菜的店面就gg了。但是当时FB的面试准备资料里share的那个练习网站真的是神器，https://brilliant.org/ 网址是这个，面过FB的小伙伴应该都知道。我当时一度迷上这个网站了，在里面做题做了好久，遇到自己不会的题或者一些比较经典的题就在纸上记录下来，前前后后做了多少道肯定记不清了，但是笔记上有大概50道。除此之外我也会看地里的统计部分的面经，遇到自己不会的知识点就上网找，一遍都能有答案，然后会吧问题和答案也记在同一个地方。

之前提到的Datacamp也是一个很重要的资源，里面有一些关于统计分布，experimentation方面的课程，我是从那些课程里第一次系统地学习并练习了了泊松分布，指数部分，permutation和bootstrap的。

最后提一个courseera上的课，也是地里的小伙伴分享的，我觉得能很清楚地帮我们理解一些看似很基础但其实很重要而大多数人不清楚的统计概念，比如random sampling和random assignment，violate each对结果的影响是什么。
https://www.coursera.org/learn/probability-intro/home/welcome

（6）Machine Learning：
这个部分地里的推荐都挺好的，我自己只上过Andrew Ng的courseera和sebastian的udacity。除此之外还上了datacamp里关于xgboost的专题讲解。我觉得Analytics Track的面试，有这些应该就够用了。

（7）Algorithm：
我没在这个部分上花太多的时间，只刷了leetcode上最高频的50道easy和medium level的题三遍，没想到居然又不止一次在面试中遇到过，也是走运。这波找工完了之后我接下来应该会找时间系统地学习一下数据结构和算法然后好好刷刷题，以后希望能做个growth engineer。

准备的材料就是这些了，接下来想说一说我实际在面试中发现自己做的好的地方以及暴露出来的一些问题，这里会涉及三个方面：behavior，product & case question以及take home challenge
1. Behavior：

这个部分是一个容易被忽视的部分，但是如果你连自己都聊不清楚，其实会让面试官，尤其是hiring manger对敢不敢用你挺犹豫的。我在这块绝大多数时候表现的是非常不错的，方法就是找几个大的topic，每个topic下面准备好故事，把故事写出来，不断地思考细节，不断地思考怎么少说废话，不断地思考怎么条理清楚。我准备的大topic有以下几个：

（1）Leadership and how to influence others
（2）A hard challenge faced and How to solve it
（3）A true failure and how to turn it around
（4）A proud success made with team together

这里重中之重的关键是少说废话，有一个behavior 很好的模板叫S（Situation）.T（Task）.A（Action）.R（Result）可以用来frame几乎所有的behaviroral 和culture fit的素材。在准备的时候一定要强调你做了什么，如果你能够量化结果的花那就更优秀了。

2. Product & Case Question
讲这个话题前要先吹一波Facebook，他家对define product & case question的定义和分类让我觉得非常make sense。在我看来IT界（不是咨询界）所有的product和case question到最后都可以被归纳到fb的两轮product 面试之下： Product Interpretation和Applied Data。

第一个内容的最终落脚点一般都是find a metrics to evaluate XXX。这个要求我们明白产品的用户，用户的问题，产品如何帮助用户解决问题，进而明确用户的goal，公司的goal，最后作为DS，我们的任务是找到metrics去quantify这些goal。 每一个公司，因为业务模式不同，最后都会一个独特但唯一的north star metrics。在面试之前，想清楚这个metrics是什么和为什么是这个在我看来是很重要的。在面试之中，当我们clarify了scope和ambiguous term之后，也应该按照步骤一步一步地和面试官讨论，把问题，产品的solution，goal这些东西都一步一步地聊出来。有的人建议先confirm goal，但是我觉得goal是在你和面试官都align了问题和产品后才能聊得出来的东西，这个大家如果有不同意见欢迎讨论。但是总结来说，这个部分的产品题，需要我们花时间去了解产品，然后一步步地去聊出面试官问你的问题的context。

在选择metrcis的时候，一定要清楚地描述分子分母，你的unit of diversion是什么，你aggregate 的time frame是什么。另外要注意的是，metrics分为三种，short term metrics， long term metrics 和 counter metrics。第一种的特点是见效快但是描述的记过不够核心，第二个的特点就正好反过来了，比如FB的CTR就是ST， retention rate 就是LT。 counter metrics是为了描述一些你不愿意看到的负向变化的。比如在FB feed里放更多的视频，你的time spend可能长了，但是你的engagement可能就会下降，因为视频是passive consume的产品，你很喜欢未必会点赞或者评论。

关于appiled data，楼主自己其实也做得不太好，fb得onsite加面的就是这一轮。这里主要就说说自己的理解和遇到的问题。 这一大类的问法都是what data would you use to XXX（我在后面会沿用同样的格式），让你brainstrom用什么data去解决问题，也就是考察在实际工作中operationalize data的能力。这里可以考察的点有很多，我争取每一个自己能想到的点都举一个我自己面试的一个实例出来供大家讨论：

（1） what data would you use to 描述impact？
e.g 某平台上突然在某一个时间点上有人说出现了很多的fake news，现在让我很短的时间出一个给VP level的报告用来描述该事件的影响。
楼主在被问到这个题的第一反应是VP level的人想要care什么impact，然后就会去想这个平台care什么impact，就去套top line metrics 比如说engagement和retention。后来面试完才发现这里漏了一个点：事件本身的影响范围究竟有多大？ 这个平台上有多少fakenews 在被产生？有多少个view是fake news？ 有多少用户看到了fake news？ 后来得到的结论是，当面对影响类的问题时，在你描述它引发的问题之前，你的第一责任应该是描述问题的scope，或者引用UX Designer经常会问的一个问题：先要搞清楚这是不是一个问题。

（2）what data would you use to signal something/find something?
e.g 某平台希望你找出business travller
回答这一类问题，我觉得先要做一些功课：想想这个产品有什么类型的第一方数据，每一个类型的数据下面有可能有什么数据，比如这个例子里，我觉得我们能获得的有用户数据，用户关系数据，用户产品使用数据，以及在使用产品时留下的源数据（metadata：e.g device，ip，gps etc）。然后有时间的话，我会再想想我能怎么吧某个数据汇总，或者吧多个数据联系起来产生某种信息，这样就多了一些derivative data points。遇到类似问题的时候就可以了调用你的信息库了。还有一个比较有用的思路是我们除了找 ‘肯定能证实的信息’ ，也可以找 ‘肯定能证伪的信息’ 。
说回这个具体的问题，首先因为是business traveller，当然就要有job。 然后关于travel，我当时选择的是先找出用户的根据地，这个可以用用户信息中的地址和用户访问源数据的最经常访问的gps和ip来结合定位。然后用gps和ip确定用户在距离足够长的地方登陆的频率，根据percentile的threshold来判断。面试完发现，这个答案是有问题的。首先，我们对相当一部分用户收集不到他们的gps和ip信息，其次我们判断的方式完全没有validation，最后我们利用的信息太少了。所以我个人的结论是：对于这类问题，不要只用analytics的方式去解决，应该要有ML， 应该要花时间去validate并人工label一些数据，然后invovle更多的feature。

（3）what data would you use to find the reason behind a increase & decrease of a certain metrics ?
e.g 某手机应用商店发现某日的应用下载量下降了，怎么找原因？
这个问题在cracking the pm interview 里有，地里的小伙伴也总结过不止一次。今天我想基于我看到的所有对这类问题的解法给一个自己认为比较全面的解：
1. 在解决这类问题前要先明白，数据在这里能提供的帮助大概率不是提供最后能用来和你的同事&面试官讨论的结论，而且提供让你们找到结论的context
2.我们需要了解以下几个方面的“context”：
2.1 trend：sudden change or gradual change ?
seasonality ? if so， maybe its normal
any special event happened internally or externally ?（new PR， new launch，system outage， new marekting campaign from competitors）
2.2 breakdown the target metrics：
这个问题里面，download是一个funnel的结果，在download之前，有访问，点击，下载，下载完成这几个步骤，每一个步骤都有绝对的量和转化率两个数字需要关注。还有一个metrics本身就是ratio，那么就要从分子分母两个方面做类似于funnel的拆解。
2.3 analyze segments：
by country， by OS， by OSV，by desktop/mobile/，etc.
每分析一个segment，需要关注的点有两个，一个是这个变化是发生在一个segment option上的，还是全options都在变化。另一个是不同segment之间的比例有没有变化，这里涉及的就是confounding和simpton paradox的问题了。
3. 很多人做完2就结束了，我觉得当我们获得了足够的context之后，应该要和面试官再聊一下基于这些context，我们应该去找谁validate什么assumption。

3. Data Challenge
关于这个部分，我觉得这个帖子已经说得非常到位了：https://www.1point3acres.com/bbs/thread-326201-1-1.html ，非常感谢这位战友分享的经验。楼主在这里就只是加一点个人的经验总结，希望对大家有帮助。如果让我用一句话总结，analytics的Data Challenge该怎么做的话，应该是：

Do as much as analysis as you can，but only showcase the most valuable findings - in a framed way。

这里有三个层面的意思：

以下内容需要积分高于 150 您已经可以浏览

（1）天马行空地去brainstrom，从最直接能想到的点去分析，到开始尝试一些需要思考才能想到的点，想到什么就分析什么，看看数据会不会带给你惊喜。
（2）判卷子的人最在乎的是你这一通分析对别人的价值，不太在乎你做的多辛苦，所以不要吧自己熬了几个晚上做的所有东西都写进报告里。报告里的东西越少越好，但你要在你的分析中找出那些是对解决问题最后价值的，按次序有选择地showcase你的deliverable。
（3）如何frame solution？ 我的看法是：describe图表-->总结出insights-->给出recommendation。describution，insight， recommendation是一个完整的逻辑闭环，它能帮助批卷子的人很快地明白了发现了什么，总结出了什么，并且依据你的总结准备建议出什么。

. From 1point 3acres bbs

以上就是自己准备面试的一些方法和心得，希望对大家有帮助。 在这里感谢一下地里的各位大神和我身边的很多小伙伴在我面试过程中提供的信息上，知识上和精神上的支持。最后祝大家面试顺利，早日拿到心仪的offer！

补充内容 (2019-3-22 01:44):
经小伙伴提醒补一下那个不work的cracking the data challenge的帖子的链接：
https://www.1point3acres.com/bbs/thread-326201-1-1.html

google, 面试经验, 数科面经, airbnb

本主题由 admin 于 2019-5-29 02:13 设置高亮