一、数据分析师职责与工作日常

1.工具

1）SQL

SQL 是架构化查询语言（Structured Query Language）Spark SQL 、Mysql 、HQL 等都是 SQL 的“方言”

一般企业内数据分析师、运营、产品等接触到的、用的是基于 Hadoop 的 HiveSQL（简称 SQL）SQL 在互联网公司的普及程度令人发指，一般来说范围是一小部分产品、个别运营、全部数分、全部 RD；某条的配置是大部分产品、一部分运营（剩下不会写的运营会改数分给的模板）和全部数分以及全部 RD；更变态的例如网易游戏的运营 SQL 和 Python 都会点。所以学会 SQL 很必要

常用编辑器：sublime、atom、notepad++，一般推荐前两个。

我自己用 Atom，使用体验比 notepad++好在可以在左侧很方便的切换文件。

2）Excel

导出的数据是需要给产品运营研发看的、所以只查出来不够，还要对数据进行进一步加工使数据具有良好的可读性或可视化。

有时候看数据趋势、分布等等也会拉个线图、柱状图等等出来看下。

3）Xmind

对于大型问题，需要考虑符合实际情况的业务指标，因此需要运用脑图进行拆解

分析师会根据个人习惯对自己常用的数据字典进行梳理，这时也会用到 Xmind。

4） Python

实际场景中接触比较少，在纯技术组的数据分析师会经常用到。

用到过一次：研发给到产品一些放在库里的网页爬取数据，数据并没有解析过，产品给到数分希望能清理出其中想要的部分。

2. 日常工作及职责

1）对流量、收入、KPI 等进行拆解

2）分析 AB 实验、做探索性分析

3）协助运营、产品、研发解决各种疑难杂症

对运营：协助梳理所需数据的逻辑、在其不明确时给出自己的整体规划及建议

对产品：通过数据给到产品同学新思路、协助拆解不符合产品同学预期的指标、分析 ABtest 效果

对研发：协助看新上线的功能对整体指标的影响；思考实验产生的数据与实验目的是否相符，若不相符将通过各种方法（抽样、AA 实验、开反向实验等）进一步修正。

4）普及常用的公司内部工具及数据基础知识

活动怎么分析、指标该怎么看、简单的拆解方法及思路、怎么给数分提需求、给到的模板怎么用

模板：

```
select a,b,c
from (select a,b,c
        from table1) A
where data='2018-08-28' ---此处可改日期
and app_id in(111,112,113) ---此处可改 app_id
group by a
```

order by c desc
limit 100 ---此处可改需要的数据条数

二、指标拆解
例 1：DAU 下降
1、推荐是否做了改动
2、拆分新/老用户
新用户下降可继续拆分新用户渠道——单独渠道下降查投放素材、全渠道掉查新用户登录是否有 bug\针对新用户的活动结束 DAU 恢复
老用户掉——查看 push 情况（进一步可拆分覆盖率、到达率、点击率 覆盖率、到达率有问题找研发，点击率有问题可进一步拆分 push 种类、push 配图大小、是否有振动、手机机型）
3.其余拆分 APP 版本、拆分国家/地域、今天有版本升级或重构等
4.都没有问题会看爱奇艺\优酷\腾讯等视频网站的用户使用时长
例 2：次留下降
1、推荐是否做了改动
2、拆分新/老用户
新用户下降会特别的看一下用户画像\用户构成（有可能是投放来的用户与产品定位存在差异因此次留偏低）
3、拆分子和分母定位问题（因为次留是个比值，因此分子减小分母不变或分子不变分母增大都会导致次留下降）、同时针对性的看互动指标（评论、播放、点赞、投稿等）
4、其余拆分 APP 版本、拆分国家/地域、今天有版本升级或重构等
5、都没有问题会看爱奇艺\优酷\腾讯等视频网站的用户使用时长

三、怎么与数据分析师沟通
1.基础知识及常见反例
1）SQL 查数据是很慢的（少则三五分钟、多则一两天都是有可能的），速度与查找条件是否复杂、使用的表的数量、涉及到的表之间关系的复杂度、分析师 SQL 能力、公司同时运行的任务个数等很多因素有关，而分析师是不能预知具体要多长时间的（但是可以根据经验判断是很快还是要等很久）。
所以如果分析师说会等很久的时候不要再问他具体要多久，他真的不知道。
也不要在自己不懂 SQL 的时候“十分体贴”的减少分析师的工作量
反例 1：
产品提完需求后说：“我只要粉丝数量排在前 100 的数据就好，是不是能写的快点？”
后果：并不会减少任何工作量，反而需要在原本的 SQL 最后加上
order by fans_num desc
limit 100
2）不要质疑数据分析师的专业性
虽然分析师干的活看起来很杂，但目前互联网公司的数据分析师统计\数学\信管\电商专业出身是标配，抛开专业讲数据组内部会有统一的分析方法，所以给到的结论会保守中立。
反例 2：
在分析出 ABtest 结果后，产品说：“我觉得结论应该是正向的，不是随机波动。”
后果：其实没什么后果，分析师一般会说建议延长观察周期或增量，但心里会 diss 你。
3）“我想看看这个数”类的需求分析师一般不接
除非是产品面对的是冷启动或类似没有头绪的问题，不然在数据组内默认不会接“想看看”

类的需求。
最好的做法是：想清楚自己看这个数的目的、预期、后续行动再与分析师沟通，明确思路的人哪怕给到的预期是“我提出的问题可能不是问题”，分析师也会愿意去一起研究这个问题给到分析。
2.沟通步骤
1）明确问题背景、意义
为什么做这件事、之前做过什么、目前的进度
2）明确需求
提出你需要分析师做什么、明确如果分析师遇到业务问题可以求助谁
1）需求沟通及确定排期
与分析师沟通确定需求、排期规划（一般 3-5 个工作日）
如果想提前排期，目前见到过最好的邮件说法：根据沟通该需求排期为 11 月 28 日，如果今天能够给到结果，这将会给我们提供很大的帮助。
2）需求交付
根据分析师给到的结果可以进一步沟通或提出后续需求。
3）结果反馈
对需求最终的结果进行反馈。
分析师一般手里会同时有 5-10 个需求，作为好的产品\运营如果及时将结果同步给分析师帮助他成长，对大家来说都是成长，分析师也会更愿意与这样的同学一起做事情。

四、如何自学
1.SQL
SQL 是架构化查询语言（Structured Query Language）Spark SQL Mysql HQL 都是 SQL 的“方言”
一般企业内数据分析师、运营、产品等接触到的、用的是基于 Hadoop 的 HiveSQL（简称 SQL）
基础：http://www.w3school.com.cn/sql/index.asp
牛客网 SQL 练习题（这里的题是 SQL SERVER 的，基本上和 SQL 不会差很多，可以用来刷）
百度“SQL 面试题”
基础的数据结构：什么是主键、什么维度、有什么数据类型等（知乎/公众号：猴子聊数据）
2.书
《excel 图表数据之道》《企业贤内助》《统计学》-贾俊平

3.网课
这个太多了，Coursera、网易云课堂很多免费的，CSDN 上很多整理的很好的博客、付费的 CDA（他家太贵）及天善智能（这个性价比还 OK）
4.证书：
小白：阿里云系列的证书（跟健旭老师走）
有统计学基础的：CDA 数据分析师等级证书（一级：业务分析师、二级：建模分析师/大数据分析师、三级：数据科学家）通过率一级在 60%-70%，每年两次（六月底和十二月底）
https://exam.cda.cn/cda_kaoshi.php?ac=kaoshi_js
5.比赛
天池、英语好的可以考虑 Kaggle、还有很多互联网大厂比如京东、美团等都会有每年一度的比赛，不过难度较大。

五、工作实例

1.用户调研

背景：调查各 APP 在性能上的用户满意程度，定位不同 APP、不同国家出现的问题

数据收集：站内信发放调查问卷

数据清洗：由于站内信发放问卷不宜过长因此无法设计逻辑陷阱题排除不干净的数据，因此清洗方法为删除未完成问卷、删除答案前后矛盾的问卷、删除问卷选择机型与收集到的用户机型不符的问卷。

数据分析：分为定性、定量两部分。定性主要看用户打出的评分，定量看选择对应评分下用户在该情况下真正的用时。通过相同 APP 相同国家不同问题、相同问题相同国家不同 APP、相同问题相同 APP 不同国家三个方面做对比，得出结论

结论举例：印度用户在开机显示视频图片时长问题上不满意（分数：3.4），与其他国家的 gap 值为 0.6。

3. 审核

背景：举报功能分为用户举报、视频举报、评论举报；每部分的举报原因均不相同，产品希望了解各举报原因频次从而进行举报原因优化。【此处有个举报频次的分布图】

数据分析师需要有自己的建议：

1. “其他”过多是否说明有原因并未被列出
2. 频词较低的具体原因是否应该下架
3. 通过数据对比各国举报原因分布，是否存在翻译后的语义存在模糊的现象
4. 极值是如何产生的，是否存在一人举报过多的“不正常”行为，是否需要限制举报频次，如果需要应该给到哪些其他数据进一步优化。

六、数据分析师简历及面试常见问题

问题 1：你认为数据分析师的职责是什么？\数据分析师是干什么的?

问题 2：请说一个通过数据分析得到结论及结论应用后的收益的实例。

问题 3：如果让你分析微信朋友圈中上线广告这个功能的收益，你会考虑哪些维度?

问题 4：小视频在达到一定的播放量后进入审核，现在需要确定进入审核的播放量的阈值，你该如何确定?

问题 5：估算每个月北京市卖出的油条数量

问题 6：你认为数据分析师需要具备哪些能力?

问题 7：数据分析师的职业规划