完整的数据分析主要包括了六大步骤，它们依次为：分析设计、数据收集、数据处理、数据分析、数据展现、报告撰写等。

1、明确分析目的

     明确数据分析的目的，才能确保数据分析有效进行，为数据的采集、处理、分析提供清晰的指引方向。

2、数据收集

     数据收集按照确定的数据分析的目的来收集相关数据的过程，为数据分析提供依据。一般数据来源于数据库、互联网、市场调查、公开出版物。

3、数据处理

     数据处理包括：数据采集、数据分组、数据组织、数据计算、数据存储、数据检索、数据排序。

4、数据分析

     数据分析分为：定性数据分析是指对词语、照片、观察结果之类的非数值型数据进行的分析。验证性数据分析是侧重于已有假设的证实或证伪。探索性数据分析是对数据进行分析从而检验假设值的形成方式，侧重于数据之中发现新的特征。

5、数据展示

     常用柱形图、饼图、折线图等图标展示有用的信息，一目了然的发现数据的本质与作用。

6、报告撰写

     报告撰写是整个数据分析的最后一步，是对整个数据分析过程的总结。一份优秀的报告需要一个名确的主题、清晰的目录、图文并茂描述数据、结论与建议。

2、假如我是广告主，你是微博的数据分析师。我现在找微博投鞋子的广告，给了200w，2星期用完。2星期后需要你给我一份广告的效果评估报告，你会怎么分析？

完整的体系框架，然后能有具体的指标来衡量。  
1、关注网站流量  
一般关注网站的IP、PV、UV等数据，这也是网络营销推广的一个常规数据来源统计。

PV(访问量)、UV([独立访客](https://www.baidu.com/s?wd=%E7%8B%AC%E7%AB%8B%E8%AE%BF%E5%AE%A2&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao" \t "https://zhidao.baidu.com/question/_blank))、IP（Internet Protocol）

2、关注页面转化数量转化率=成交行为的人数/进店的全部人数  
3、页面的停留时间

4、被关注的趋势

5、销量、利润变化

6、广告的展现量：广告在对应的广告平台上面的展现次数，俗称曝光量，这个越大接触的用户越多，对后面的转化率的提高就越多。  
7、广告的互动量：互动量体现在用户是否参与了评论、转发或者点赞，当让如果有坏的评论也可以接受进行下次的优化，  
8、广告的转化量：转化才可以算是广告生命周期末尾，具体就是用户看到你的广告买了产品、下载什么软件、报名什么课程等等

1.sql题：  
已知一个订单表，有字段【订单号】，【日期】，【下单时间】，【下单金额】，【商品一级类目】，【商品二级类目】  
请求最近7天内每一个一级类目下成交总额排名前3的二级类目。说一下思路：

1.先筛时间7天内,datediff(now(),日期)<=7

2.分组group by一级类目

3.sum开窗求每个二级类目销售总金额，sum(下单金额) over(partition by二级类目) as total\_amt

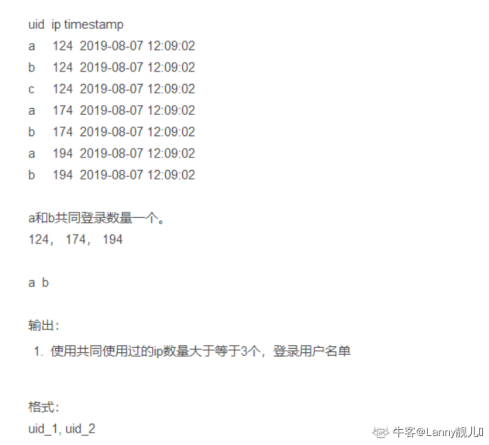
4.再用rank(total\_amt) over() rk rk in[1,2,3]或order by total\_amt 结合Limit 3选前3。

分析和挖掘你更偏向哪个方向？

（1）你的一个朋友经营一个网店，以前100个人浏览会有50个人下单，最近客服咨询量提升到200，但下单却下降到40，请你帮她分析原因。  
答：先锁定原因，是在咨询到下单过程中，所以可能是页面设计的不合理，支付体验不佳，或者同类商品推荐更具诱惑，考虑商品问题。好像都没回答到他想听到的点，又问还有吗，然后我说也可能是客服的原因，他才接下来第二个问题  
（2）你的朋友说最近咨询的人变多了，但是他做的是小本生意，没有能力再去请多一个客服，所以导致自己回复用户速度下降，很多人就走了没有下单，请你帮他想一下解决方案  
答：能请多一个人更好，不能的话就看一看哪些时间段成交率比较高，在这些时段提高自己的效率，其他时间段可以稍微放松。或者重点服务信誉高，成交率高的用户。然后面试官又问还有吗，我实在想不到，沉默了一会....后来面试官说可以设置自动回复。  
（3）你的朋友按照你的建议去做，生意变好了，他赚钱了，然后他一下子请了5个客服，请你设计指标去考核客服的KPI。  
说了一些成交率，回复率，服务态度等，想不到别的啥。

讲了一个实习的项目，问的提别细，每个细节都问的特别清楚，我做了什么工作，怎么做的，具体实现。

1. 做个sql题。意思就是有些人一个人开了好几个小号，你要用ip登录次数识别出来。大家可以讨论下，我没做出来正确方法，但最后问了面试官怎么做。



1. 有没有用过拼多多，怎么评价砍价活动的广告效果？有哪些指标？
2. 某个平台618大促后销量下降了，分析原因。
3. 针对某平台的销售数据，每家销售店铺有每日成交额，找出成交额在六月份前20%，20-40%，40-60%，60-80%，80-100%的店铺；
4. 涉及的指标

3.sql：一个用户登陆表，里面有uid,log\_page,log\_time，log\_page就是A、B、C、D、E....网页,log\_time就是对应网页的登陆时间，找出所有的登陆过A、...B、...(非C)...D的用户，就是先登陆A网页，之后登陆B网页，之间可以登陆其他所有网页，再登陆D网页，B和D网页之间可以登陆任何非C网页的其他网页。这个题目没答出来，面试的小哥哥很好，一直引导我，感觉得花了半个小时，结论是用left join，先找出用户A的时间，然后left join到所有A之后的B，再找到所有B之后的D，根据此表，对应一个C网页的时间表，只要用户C网页时间在任意B和D之间没有就输出。（不知道讲清楚没有~）

4.也是sql：对一张用户的订单表，有下单时间，统计出上个月每周都有销量的uid数量：我说的是将表切割成每周，然后inner join每个切割表中都存在的uid，面试官说太复杂，直接判断，每个周都存在购买就可以，后来想可以用case when来对每个周有购买的做记录，最后以相加大于4就可以。

5.sql中where和having区别；

5.场景题目：如果一个购物平台页面，两个子入口连接对应的产品列表，交易额相差很大，分析原因：可以从商品类别差异、产品销售量、单价、入口流量、各渠道转化率来分析（后面是面试官补充的）。

1. 一个应井盖制造商，制造方形和圆形的应井盖，现在为一个城市定制应井盖，怎么决策该制造方形还是圆形？后续如何决定二者的生产比例？

1、你的专业和应聘岗位差距这么大，你是如何自学的？在自学过程中最有效的途径是什么？

2.目前拿到的offer和面试到达的轮次

3.家在哪里？是否是独生子女？以后期望工作的城市？

4.你和科班出生的相比，你的技术是领先还是落后？

5.对你打击最大的一件事和最引以为傲的一件事

6.你和你的leader意见冲突，如何解决？

7.你最大的一个缺点

8.你的兴趣爱好

9.为什么没有考虑实习转正

2.两端实习经历，哪一段收获更大？

3.挑选一个最喜欢的项目介绍过程

4.算法题：求子数组的最大和

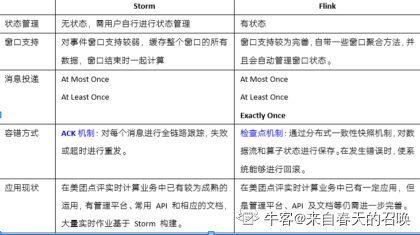
5.hive内部表和外部表的区别，各种join解释一下，sort by和order by的区别

6.决策树如何分裂节点，三种决策树的差别

7.使用拼多多感受如何？如何决策是否要上线购物车这个功能？如果不用ab test，仅用历史数据又如何决策？  
拼多多是自己提前批第一个走完全流程的，感觉就是临场还是会被问到许多不会的点，比如sort by，自己写hive几乎没用过，猛然被一问就懵逼了，简历上的点都要搞懂，然后多发散问自己几个为什么，希望大家都能收获自己心水的offer！

1. 刚开始面试就是对简历项目一顿撕，PV和UV是怎么计算的，UV怎么进行去重的？不用ES该如何实现去重？

2. 说说flink,spark streaming,storm的区别？



3. spark的调度执行逻辑，stage，宽依赖和窄依赖，容错机制？

容错机制：

窄依赖可以通过血缘关系 来恢复故障RDD，而宽依赖则考虑使用 检查点 的方式恢复。

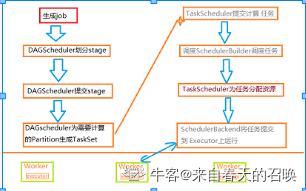
RDD的容错机制是 如何实现的？

1. 借助这些依赖关系，DAG可以认为这些RDD之间形成了 lineage(血统，血缘关系)，借助lineage ，能保证一个RDD在计算前，它的父RDD已经完成了计算，并实现了RDD的容错。

2. 当某个RDD发生故障需要恢复时，从数据源逐步执行对应的transformation操作来重新生成当前的RDD，这种容错策略 要高于 常用的 检查点 （Check Pointing）策略。

CheckPoint(将RDD持久化到磁盘上):

调度执行逻辑：



作者：来自春天的召唤  
链接：[https://www.nowcoder.com/discuss/226834](https://www.nowcoder.com/discuss/226834" \t "_blank)  
来源：牛客网

依赖关系：

依赖 关系：

1.  窄依赖【Narrow Dependency 】（1:1）：如 map ->flatMap等。

2. 宽依赖【Wide Dependency 】（1：多）：如groupByKey()等。

Shuffle：

存在spark shuffle :

因为具有某种共同的特征的一类数据需要最终汇聚 （aggregate）到一个计算节点进行计算 ，这个数据重新打乱然后汇聚到不同节点的过程就是 shuffle。

老 ：

Hash Base shuffle 产生的临时文件数 = MapTask \* ResultTask 。

弊：

会产生过多的临时文件。

新：

SortBasedShuffle： 产生的文件数 = MapTask 数量 。

如果Shuffle 不落地：

①可能会造成内存溢出

②当某分区丢失时，会重新计算所有父分区数据

Stage:

一个DAG会根据RDD之间的依赖关系进行Stage划分，流程是：以Action为基准，向前回溯，遇到宽依赖，就形成一个Stage。遇到窄依赖，则执行流水线优化（将多个连续的窄依赖放到一起执行）

1. 分布式一致性协议算法有哪些，说说它们？
2. 说说一致性哈希算法？

7. 说一下布隆过滤器（Bloom Filter）？（自问知道的比较清楚，说了一点，我等他往更深层问呢，结果他直接就下一个问题了，早知道我就自己说完了。[是谁说阿里考官问知识点直到不会为止的，出来我要跟你理论理论]）

场景：快速 从海量数据中找出某个成员是否属于这个集合：

10. Mysql问的比较少，只问了索引的数据结构？

答：索引的数据结构：

1.生成索引，建立二叉查找树

2.生成索引，建立B-Tree

3.生成索引，建立B+-Tree

4.生成索引，建立Hash，基于InnoDB和MyISAM的Mysql不显示支持哈希

5. 位图数据结构，BitMap,少量主流数据库使用，如Oracle，mysql不支持，

11. volatile和synchronized关键字？

Synchronized 与volatile 关键字的区别？

概念上的区别：执行控制 和 内存可见；

执行控制：控制代码执行顺序以及是否可以并发 ；

Synchronized 解决执行控制问题，它会其它线程获取当前对象的监控锁，使得当前对象中被Synchronized关键字保护的代码块无法被并发执行，并且Synchronized 还会创建一个内存屏障，保证了所有CPU操作结果都会直接刷到主存中去，从而保证了可见性。

2.为什么觉得自己逻辑能力和学习能力强，举例论证  
3.你认为机器学习在数据分析中起到什么样的作用  
4.聊了一下最近一次实习的项目  
5.项目中的特征怎么选取的  
6.讲一下假设检验（这里我展开太多了，写了各个统计量的计算公式，不是面试官想要的答案）  
7.假设检验的应用场景  
8.因果和相关的区别  
9.一个天平和七个砝码，其中五个一样重，另外两个稍轻，问把这两个轻的找出来，最快需要称几次

（这个问题回来跟同学商量了一下，暂时还没有找到公认的最佳方法，同学提到的方法如下，供参考，欢迎讨论）

砝码分成3-3-1，三个三个称一次。

1.如果两边一样重，则两边各有一个轻的，每三个中再挑两个称，找到答案。共3次。

2.如果两边砝码不一样重，挑出轻的那一份，从中选取两个进行比较。

①如果两个一样重，拿出第一次称的重的那3个中的一个，与这两个中的一个比较。

如果一样重，则剩下没称的2个是轻的；

如果不一样重，那么刚刚称的一样重的两个是轻的。共3次。

②如果不一样重，则找出一个轻的；再称一下剩下的两个谁轻谁重，得出答案。共3次。

内存可见：控制的是线程执行结果在内存中对其它线程的可见性，根据Java内存模型的实现，Java线程在具体执行时，会先拷贝主存中的数据 到线程本地（CPU缓存），操作完成后再把结果从线程刷到主存中；

volatile 解决了内存可见，所有对volatile关键字的读写都会直接刷到主存中去，保证了变量的可见性，适用于对变量可见性有要求而对读取顺序没有要求的场景。

两者区别：

1. volatile 不会造成线程的阻塞，Synchronized 会造成线程的阻塞。

2. volatile仅能使用在变量级别，Synchronized 可以使用在变量，方法，类级别上。

3. volatile 标记的变量不会被编译器优化，Synchronized标记的变量会被编译器优化。

4. Volatile 仅能保证变量的修改可见性，不能保证原子性，而Synchronized 可以保证变量修改的可见性和原子性。

5. Volatile本质告诉JVM 当前变量在寄存器（工作内存）中的值是不确定的，需要从主存中去读取 ，Synchronized则是锁定当前变量，只有当前线程可以访问该变量，其它线程不可以。

12.高并发问了一点，ArrayBlockingQueue和LinkedBlockingQueue?

13.最后还问了一个IPV4相关的场景，设计算法和数据结构，晚上面的，脑子有点闷，没想明白？

14、猪死了，为什么。

回来想了一下，应该先问时间、地点、猪的数量、周围环境，如果分析得当，5个左右的问题就可以问到关键点  
  
15、数据挖掘项目介绍（这里追问的比较细，包括为什么这样定目标，为什么要做分类，样本是什么，样本量多少，为什么不采用另一种方式取特征，预测的时间范围，最后效果如何，如何在业务上验证有效而不是在验证集上。）

作者：-offer会有的！-  
链接：[https://www.nowcoder.com/discuss/226713](https://www.nowcoder.com/discuss/226713" \t "_blank)  
来源：牛客网  
  
3. 另一个项目，涉及综合评价，PCA，FA。问PCA从数据到最后结果的流程，每一步数据如何处理，选取了几个主成分，怎么确定的主成分数量，怎样从主成分到综合分数。  
4. 写了个SQL。订单表，有ord\_id,mall\_id,goods\_id,sale\_number,amount,求每个商店里的商品价格的中位数。（用到了开窗函数和join，我大体方向正确，但是考虑不周全，面试官有引导纠正）  
5. MySQL数据库，表的主键只能是一列，还是可以多列？（我理解成了可以有几个主键，面试官想问的是主键能否是列的组合）  
6. 平时主要用Python做什么，会做策略和算法吗  
7. gbdt和xgboost的区别（有追问基学习器，为什么gbdt用的是树，我提到了CART树，面试官问可以用别的树吗，这个我不知道……还有gbdt可以用线性的基学习器吗？为什么xgboost可以用线性的基学习器？这个我也直说了不知道……求大神解答）  
8. 对拼多多的理解，对拼多多产品提一提看法（我提到的朋友分享的砍价页面  
9. 砍价页面，不同的用户点进来，每个用户每一刀砍掉多少钱，你怎么设计，写个函数吧（我说了想法，没写）  
10. 你怎么验证刚刚提到的砍价策略是有效的，需要哪些数据指标（我提到了砍完价之后的成交率）  
11. 砍价的成交率越高越好还是越低越好（这里发现了我对拼多多业务理解的盲点，成交率越低越好，因为每成交一单都会增加平台的成本，这里边只有成本，没有利润空间，拼多多主要赚的不是商品利润，而是流量）

1. 最后做了一个sql，是母婴类花王牌的好评率，这个题之前看到过，好像是以前的真题，也不难。

2.实习经历介绍

3.对埋点的理解

4.对数据分析这个岗位的理解

5.说说实习的时候所做过的比较满意的案例，然后追问（我说了用聚类做用户分群，然后就问了kmeans的原理，步骤，距离的计算方式）

6.问还了解其他算法模型不?问了朴素贝叶斯

7.说一个你接过的做的比较好的临时数据需求

8.数据分析的流程

9.你想做哪方面的数据分析工作

1、上来自我介绍，简单讲了两个项目、

为什么选择数据分析师、对数据分析了解多少？

spl：班级成绩前两名的学生信息；

概率题；54张扑克牌，平均分成两份，这两份都含有两张A的概率；

业务场景题：有一款app、在某一个月里，日活率逐渐下降，请做一下分析；

说一下GBDT原理；

说一下pca原理；

面试官：我觉得你应该去投算法岗，为什么选择数据分析呢？

我：我爱数据分析，我对数据分析有浓厚的兴趣！！！（内心：算法岗太难了）1.自我介绍

2.数据分析师和算法工程师的区别

3.用代驾这个业务构建一个用户画像，期望得到什么结论

4.介绍项目中的算法，如kmeans，dbscan，谱聚类，自编码网络，xgb等

5.做两道题，一个sql，一个python

6.会使用tableau，ppt吗

7.有什么要问的

1. 自我介绍 教育背景+项目经理+实习经历

2. 之前在学校做过的用户画像和在公司做的用户画像详细说明一下，小姐姐问有了解过其他地方是怎么做用户画像的吗？

3. 实习都用的是啥查询语言，答用过mysql、hive和presto，问那你说一下mysql和hive的区别吧.......然后我就给自己挖了个坑，最后我说了句hive不支持不等值连接，mysql支持

4. 问 那正好问下你怎么实现两个1亿的表做hive不等值连接且避免数据倾斜的情况，然后我就把自己能说的都说了。

后面就是我问小姐姐问题了，顺丰内部这个岗好像分3个方面，偏重不同的方面。

二面：昨晚收到的面试通知，实习偷偷溜出去半小时面的，本来说好是视频面试，但因为开了三次都是要么黑屏要么听不到声音，就电话面了

1. 万年的自我介绍

2. 说一下在之前BI数据分析实习部门做的成果-例如dashboard可视化成果，得出了什么结论且给出了什么建议；

3. 目前在的数据分析部门主要做什么，业务内容以及工作内容；

4. 根据我的实习问了，资源位设置的合理性给予数据支持举个例子以及曾经做a/btest的案例；

5. 问了个sql，就是涉及开窗函数-分组排序；

6. 开放题：如果客诉率在一段时间内升高，你需要在下周一给大老板一定的分析结果，你会从哪些方面考虑。（感觉我这个在瞎说......小哥哥还一直说 你还有什么需要补充的嘛？我说不下去了，主要想到的是首先拆分出指标标签，再可能运用聚类方法，找出大类，后面还说了个有客诉记录的话，可以与语义分析结合搞）  
一道sql题目 如何找出每一个级别下面购买最多的十个用户  
应该用窗口函数的，但我忘了，用了个group by 错了  
  
一道业务题，AB TEST  
  
一道概率题，三变量贝叶斯公式，就写了一下思路，不知道对不对

2.首先问了一下简历里建模用到的最小二乘法，及为什么要那样求解（这个应该还算简单）。

3.A/B test，怎么做，什么时候用z test，什么时候用t test。

4.然后是SQL题。

①面试官看我笔试第3题没写，让我现场重新写一下，之前有很多牛友发过了笔经，不再重述。就是求新增用户数、次日留存率、30日留存率。

②表order有三个字段，店铺ID，订单时间，订单金额，查询一个月内每周都有销量的店铺。

这两题第一题笔试后我有思考，所以顺利的写出了新增用户数，但是面试的时候脑子抽了，求留存率我只计算了留存的人数忘记除了，面试官提醒了一遍没反应过来。第二题个人感觉难度比较大，面试时压力大思路混乱，用了lead()over()开窗，但是最后判断有一周没有销量的店铺ID的条件错误了。可能会有更好的查询方法。后来想了下，应该令销量等于0，然后查询出如果有连续7天及以上的日期的店铺ID，最后排除这些ID。

5.业务题：问如何预估用户的价值，用什么算法预估，用什么算法计算价值。答:根据用户的购买力（问什么是购买力，自己给了个定义）将用户分类，用决策树算法分类，用线性回归方法计算用户价值。本人对机器学习算法不是特别了解，临场凭感觉回答。（后来问面试官，面试官并没有说回答错）。

6.有什么问题问我。我问了流程大概有几轮，大概中间隔多久，面试官说这个他不清楚要问HR。然后问了对我有什么建立，有什么需要提升。面试官接下来的回答扎心了，面试官说感觉我的SQL还不太行。。。。（本人一直最自信的技能就是SQL。。。。在SQL上下了很多功夫）。

至此结束，一共一个小时左右。

面完后就感觉不是特别妙，但觉得自己应该还有机会二面，最后没有过也有些失落。Anyway，再接再厉吧，再多练练SQL，复习概率知识。PS：感觉拼多多的数据分析面试多是根据自己个人的简历，简历上写了什么会安排相应的面试官着重问，大家要面拼多多的多注意自己的简历细节。

3、一道sql：使用共同的ip数量大于等于3个用户对

4、logistic模型的原理与求解

5、xgb和rangdomforest是同一种集成学习吗

6、有用过拼多多吗？关于砍价活动怎么看？选取指标评价砍价这个营销活动

1、自我介绍

2、写sql：取到当日支付订单数前1000的商品（订单金额>=80)，并且取到这些商品中每个商品订单量前100的渠道

我的sql是自己在网上学的，刷了一下牛客网的题，后半部分写了一下只能输出100个，面试的小哥哥耐心的引导了我好久是组内排序我也没写出来，还是自己太菜。

3、从数据分析师的角度来分析一下销量下降的原因。

答，从内部和外部，后让我着重分析内部

4、估计上海红绿灯的数量

5、你认为数据分析师的工作及三个使命

6、对拼多多的看法