> 本文来自一个8年大数据老兵的面试经历投稿，我完完整整的看了一遍，真的很细很细，疫情期间面试各种失败各种总结，最后拿到Offer实属不易，精神很值得大家学习

前言

我不是什么大牛，我只是一个有八年工作经验的互联网老兵，既没有非常亮眼的学历，也没有牛逼大厂的履历。

这个冬天，在孩子得病从急诊转住院的时候，我得到了年前将被优化的消息，作为家里唯一经济来源的我整个人蒙了，一时间茫然和无助差点毁了我。

最后我还是和家人如实说了，受到了他们的极大的鼓励，也找到了重新开始的勇气。

可惜这场疫情来的如此凶猛，职位少、要求高、电话面试、视频面试、在线coding、屡战屡败、屡败屡战，构成了我这两个月的常规生活。

我一开始也焦虑、茫然，甚至对自己的能力和工作经验深深怀疑。

后来经过几个好朋友的鼓励打气，也看了敖丙的很多面试文章，认真总结自己面试中不足的地方，查漏补缺，终于在这周确定了offer。（这是原话，我真没瞎加哈哈）

接下来我就我之前面过的腾讯、高德、京东、美团、饿了么、快手、字节跳动、滴滴、360金融、跟谁学、网联清算、华晨宝马、快看漫画、陌陌、脉脉等等等公司的面试题做一下总结，权当抛砖引玉，希望对大家有帮忙。

面试问题汇总

基础问题

linux和网络基础

（1）linux系统内核态和用户态是什么，有什么区别?

（2）BIO、NIO、AIO都是什么，有什么区别?

（3）TCP和UDP的区别?

（4）详细叙述TCP3次握手，TCP和HTTP的区别，其中字节面试官问的最细，他会具体问TCP底层的3次握手的具体实现逻辑，第三次握手如果失败会怎样。

建议把TCP关闭时的4次挥手也看看，敖丙的文章就有，看了至少表面的东西难不倒你们，由于这个是最基础的问题，如果回答不好，面试官的印象分就你懂得。

（5）rpc和http的区别，你知道有什么rpc框架。

（6）https相对http都实现了什么加密方式，是对称加密还是非对称加密?

（7）用linux命令怎么做分组求和，怎么把字符串根据分隔符变成数组（这里建议大家读读敖丙的linux命令篇）

JVM基础

（1）简要介绍一下JVM虚拟机（这个问题不是把JVM分成JMM，类加载和GC来问，一定要想好怎么描述JVM）

（2）简述一次GC的过程（Minor gc和Major gc过程还记得么）

（3）JMM是什么?

（4）JVM共享内存都有什么，什么是堆外内存?

（5）GC区域，垃圾回收算法，垃圾回收器，G1、CMS、ParNew等垃圾回收器的简介和之间的区别。

（6）类加载过程（5个过程最好能研究明白，因为还涉及到栈帧、局部表量表、操作数栈、动态链接和方法出口等知识，去看一下敖丙的文章就明白了）

（7）一个ArrayList的两个对象的getClass()得到的结果相同么（理解类加载和Class类类型）

（8）死锁怎么查问题（-XX:+PrintGCDetails）

（9）Gc日志得会看，尤其问到怎么查OOM问题的时候，你应该知道使用jconsole，jstat，jmap，jvisualvm等的工具来查看gc状态，看看是不是年轻代设置太小了导致major gc频繁或者内存泄露了。

JAVA基础和多线程基础

（1）synchronized在JDK6做了哪些优化，synchronized和lock的区别

（2）懒汉单例用duble check是线程安全的么，为什么要加volatile

（3）Volatile有什么用，什么是CAS

（4）什么是happens before原则

（5）什么是AQS

（6）线程sleep和wait的区别，线程join是什么意思

（7）Java都有哪几种锁（敖丙的文章）

（8）线程池分几种类型，其中的coreSize、maxSize、存活时间、等待队列、拒绝策略要清楚

（9）Java乐观锁的实现（CAS+自旋）

（10）阻塞队列的实现，至少自己会实现2种阻塞队列的方法（单锁，多锁， ReentrantLock，Condition）

（11）CountDownLatch、CyclicBarrier、Semaphore区别，使用场景

（12）HashMap是线程安全的么，底层怎么实现的（get，set，resize），JDK1.8之前和之后做了哪些修改，如果要使得插入kv有序需要使用哪种HashMap（LinkedHashMap，TreeMap），ConcurrentHashMap线程安全是怎么实现的（JDK1.8前后实现不同）

（13）ArrayList和LinkedList的区别，栈和队列的区别。Queue和Deque区别

（14）Netty，Jetty实现原理。

（15）Java 静态代理、动态代理

（16）Forkjoin模型

（17）Java回调

（18）协程和线程的区别

（19）JDK1.8有什么新特性，了解函数式编程么（不了解的看看guava）

数据结构算法和设计模式

（1）设计模式一般引申自项目或者工具底层实现，所以需要懂一些比较常见的设计模式，工厂、单例、观察者、命令、适配器、代理等等

（2）算法主要是查找和排序，所以至少要会手写主流的排序算法和查找算法

（3）LSM树是怎么实现的。和mysql的B+树有什么区别（LSM树是hbase和levelDB的底层存储的结构，不懂不应该）

（4）二叉树，平衡查找二叉树，红黑树等

（5）栈，数组，链表，队列，双端队列，跳跃表（redis zset）等

spring系列

（1）AOP，IOC概念

（2）Spring cloud组件介绍，具体问的比较多的是hytrix和eureka，hytrix主要问怎么实现限流和降级（线程池和信号量），两种实现方式有什么区别，具体熔断时的配置；eureka主要介绍和zookeeper的区别，以及注册流程

（3）Spring boot配置很多都注解化了，所以常用的注解要知道

（4）过滤器和Spring拦截器的区别

消息中间件AMQP

看敖丙的文章就够了

redis缓存相关

看敖丙的文章就够了

（这两段没笑死我）

其他类型

（1）单点登录系统怎么做（SSO系统）

（2）为什么选择cassandra而不是hbase，两者有什么区别

大数据问题

hadoop

（1）hadoop1.0的进程都有哪些，hdfs和mapreduce简介

（2）集群初始化的时候namenode都做了哪些工作，fsimage和editslog都是什么

（3）SecondaryNamenode有什么作用。

（4）Hadoop读文件和写文件流程

（5）Mapreduce过程简介（注意这个是基础，不会说拉低印象分），shuffle流程，jobclient提交job的流程等。

（6）Mapreduce怎么进行序列化反序列化的（inputFormat，outputFormat）

（7）Jobtracker都有哪些任务调度器

（8）Hadoop YARN都做了哪些优化，YARN都有哪些进程，YARN提交job的流程

（9）Mapreduce优化（mapjoin，combiner，小文件合并等）

（10）简述hive表join怎么用mapreduce实现，mapreduce二次排序，二次排序分区和分组的区别

（11）Hadoop集群HA实现（zookeeper实现主备和federation最好都弄懂概念）

（12）其他框架比如spark怎么和yarn集成的

（13）Spark相比mapreduce的优化（内存计算，RDD等）

（14）给你100亿条数据的用户表和一块100MB内存，怎么去重或者判断一个用户在不在其中（bitmap，布隆过滤器等）

（15）加分项：读过hadoop源码么，具体哪一段源码介绍一下。

hive

（1） Hive数据仓库的架构

（2） Hive怎么把sql转化成mapreduce的（至少知道sql解析器解析成AST语法树，后面解析成queryblock，进执行队列等等）

（3） Hive基本数据类型，组合类型（当时问Hive中的int类型有几种，蒙了）

（4） Hive底层存储类型，压缩格式

（5） Hive UDF，UDTF，UDAF，窗口函数（row_number, rank,cube,rollup,lag,lead）（一般是跟着sql coding来问的）

（6） Hive优化（count(distinct xxx),去除null值，小文件合并，map和reduce个数优化，解决数据倾斜）

（7） Hive分区和分桶的区别。分桶主要解决什么问题。内部表和外部表的区别。怎么动态分区。

（8） Hive怎么自动补全分区（MSCK命令，这个比较冷僻，知道有这个东西就行了）

（9） HIve列存储，rcfile和orcfile和parquet怎么存数据的

hbase

（1）hbase架构简介

（2）Hbase怎么读写数据详细流程

（3）Hbase的应用场景

（4）Hbase优化（热点，预分区，rowKey设计，手动合并等）

（5）Hbase为什么写快读慢（LSM树）

（6）Hbase是cp还是ap架构？（CAP理论看懂没有，hbase是CP的）

（7）Hbase 怎么scan数据的。

kafka

为什么kafka放到大数据里来说，因为kafka大部分场景下是ETL流程和流式计算流程的source端

（1）kafka架构简介

（2）Kafka为什么快，性能好，吞吐量大（mmap和sendfile了解一下）

（3）Kafka会丢数据么，kafka消息有序么

（4）Kafka producer consumer怎么实现at most once和exactly once（幂等计算和事务）

（5）Kafka 高可用怎么实现的（AR，ISR，OSR）。会不会脑裂（不会啊，参照zookeeper选举）

（6）Kafka leo（log end offset）和hw（high watermark）

（7）Kafka consumer消费topic的某一个partition时，不同group和同group中的消费者有什么不同。

（8）Kafka ack有几种，每种什么意思

（9）Kafka有什么坑，怎么改进（大脑风暴了）

（10）Kafka相比rabbitMq等传统消息队列有什么区别

zookeeper

（1）zookeeper简介

（2）Zookeeper节点类型

（3）Zookeeper watcher机制

（4）Zookeeper使用场景，怎么用zk设计主备高可用，怎么用zk实现分布式锁（看敖丙文章去，其实就是临时顺序znode的建立和watcher机制的妙用）

（5）Zookeeper选举机制，会不会脑裂

其他工具

因为我没有很多spark和flink的项目经验，所以这部分问的较少

（1）介绍一下storm，spark，flink

（2）Spark RDD

（3）Spark stage怎么划分task的

（4）Spark宽窄依赖

（5）Spark shuffle

（6）Spark 为什么容易OOM

（7）Flink 窗口类型都有哪些

（8）Flink 水位线是什么，要解决什么问题，怎么保证消息有序

（9）Flink 怎么实现exactly once

项目中涉及到的mysql、redis、flume、sqoop、es等工具也会具体问的，这里我就不详细说了，redis和mysql的可以直接看敖丙的文章就可以了。

数仓相关的问题和数据分析、算法端的问题

（1）你是怎么设计数仓的

（2）数据仓库是什么，和数据库有什么区别

（3）你的数仓怎么分层的

（4）维度建模的流程，其他类型的建模方式

（5）Inmon模型和KimBall模型有什么区别

（6）怎么提炼业务指标

（7）怎么设计事实表和维度表

（8）数据立方体的一些概念

（9）什么是缓慢变化维，怎么处理这种缓慢变化维

（10）具体项目中会问到日志或者数据是增量存还是全量存，可能会引申到拉链表，甚至让我实现拉链表（字节2面挂就是拉链表流程没有正确写出来，所以后来干脆自己在mysql上面实现了一把就懂了）

（11）会用python，R语言进行数据分析么，会用SPSS，EXCEL，tableau之类的工具么

（12）使用过什么多维查询引擎（impala，kylin，presto，druid等，如果没用过别说用过，因为可能面试官很了解的话会问的很细很底层）

（13）MPP的概念，clickhouse之类的工具使用

（14）调度系统报表系统元数据管理系统血缘分析等系统设计，标签系统设计，AI算法实现，用户画像设计等

（15）谈谈对数据中台的理解，要解决什么问题（看你的思考能力和对数据部门职能的理解）

（16）谈谈对数据治理的理解

Coding

这里只给大家提供一些遇到过的简单问题，大家应该掌握基本的查找算法、排序算法，熟练使用递归、贪心，能明白动态规划更好。

Leetcode上面的题有空再去刷，因为几千道题要花费大量的时间，对于需要准备考sql的同学，建议把牛客网上面数据库SQL实战都做一遍，理解了就差不多了。

（1）实现一个函数把两个有序的int数组结合成新的有序数组（java，遇到过2次）

（2）给a[n]数组进行全排序，找到一个组合的前一个组合，比如a[3]{[1,2,3],[1,3,2],[2,1,3],[2,3,1],[3,1,2],[3,2,1]}，给出[2,3,1,]，找到他的前序是[2,1,3]（java）

（3）给定一个正数数组arr（即数组元素全是正数），找出该数组中，两个元素相减的最大值，其中被减数的下标不小于减数的下标。即求出： maxValue = max{arr[j]-arr[i] and j >= i} （java）

（4）有8个球，其中有一个比其他7个重。给你一个天平要求2次称重就把重的那个球找出来。（智力题）

（5）求一个数组中不存在的最小正整数（java，这个好像是程序员面试指南里头的题）

（6）给定用户登录表，怎么查连续3天未登录的用户（sql）

（7）给定每天收入明细数据，怎么查每一天的历史收入总和（sql）

（8）Hive 表中有重复值，怎么查一共有多少个重复值（hql）

（9）给定注册表和登录表，用一个sql求1-7天留存（sql）

（10）实现拉链表（hql）

（11）给定电商订单表，字段为订单id（order_id）和订单组合（type_list），求这个订单组合中每种类型商品的相关商品TOP10,即求这个商品相关的商品（下单这个商品的同时也下单其他商品）下单量TOP10（hql，行转列）

（12）给定一个广告投放表ad,字段有aid（广告id）和citys（投放城市city_id集合）和城市表city_info,字段有city_id和city_name（城市名称），求具体城市名称的投放广告量TOP10。（hql，行转列）

项目

面试官考察项目经验，考察的其实不仅仅是你基础的掌握，更多的是自己对业务的理解，架构设计，自己对项目的思考。所以，除了项目中涉及到基础知识的问题，还会问到诸如你觉得项目中有哪些设计比较好，或者有哪些不合理的地方，你是怎么解决的等方面的问题。

这些问题往大了去可能是架构方面的，也可能是具体技术细节。但是只要你讲出自己的思考和解决方案，有经验的面试官会大概了解到你的技术深度、架构设计能力和解决问题能力的层次。

所以一定要找到有亮点的地方提前进行背书，要有层次的介绍项目，思考一下项目设计或者实现不完善的地方。

还有一些面试官会问到如果让你设计一个什么什么系统，你怎么设计。这种题我觉得也最好提前做过背书。

因为对于工作经验少的同学，面试官主要看他想问题的深度和广度，但是对于工作经验不少于5年的人来说，面试官更关注你是否有成熟的实现流程和方法论。

所以一个层次化流程化的设计会极大增加面试官好感度。切记避免废话连篇核心不明确。（这也是我的问题，因为没有准备，所以遇到肯定说的很散，这样面试官觉得你自己做事情没有核心和方法）

个人价值观

一般技术面到后面，面试官都是leader或者是部门老大，他们其实很关心你的职业规划、对待工作的态度、团队合作的能力、自我价值实现方面的思考，当然还有项目实现的能力，过往项目经验和深度。所以最好自己先想想怎么用简短的话表述清楚。注意围绕关键词去说。

写给看到最后同学的话

这些算是我作为过来人对于应届毕业生和刚工作不久的同学的一些小小建议吧。

（1）一定要紧跟技术前进的脚步，尤其是大数据相关的技术，在技术更迭的时候一定要学习熟悉新技术，看源码，哪怕自己在工作中用不到也一定要学。因为这是你下一份工作的敲门砖。

我作为一个老兵，在上一家公司工作4年，公司没有spark和flink的业务场景，我也没有逼迫自己学习这些新技术。

结果现在面试碰壁，其实大部分原因就是人家用的主流技术就是这些，你不会你就被淘汰。

（2）不要给自己设置舒适区，这个就是说，一个公司待久了不要懒惰，不能荒废自己，始终要保持清醒的头脑和进取心，不断学习，不断完善自己的技术，架构设计能力，项目管理能力，交付能力等。

一定要及时从项目中总结经验和不足，最好落实到日记本中，最后通过不断思考，形成自己的做事方法论。

（3）对自己的职业生涯要有一个规划，以后要做哪一块一定要有自己的想法，确定了就要从这个方向完善自己，多学多练。

目前大数据这一块，有数据中台架构的公司不是很多，除了算法岗外，大多数人在团队中都是1专多能的角色，今天干干ETL，明天搞数仓，后天又给BI出数据，可能又搞调度系统、报表系统、标签系统、反作弊平台等平台。

没有人会专一做某一块，但是自己一定要想好哪一块是自己以后要走的方向，那么这一个方向确定了就要深入的学习这一块的知识，多看源码，多做练习，如果接触到具体项目，要在项目中沉淀自己，最后形成自己的知识体系。

（4）做事情要有担当，不要根据OKR给自己设置界限，有能力有空闲多做一定要多做，这也是别人认可你的最佳途径之一。互联网圈子很小，大家认可你，以后去大厂，换个好工作，也许就更容易。

敖丙的絮絮叨叨

是的结尾我还是说一下吧，投稿的读者是一个互联网经验的老兵了，两个月的高强度面试也让他有了很多收获，有一句话我很喜欢，不要给自己设立舒适区，要有危机感，真的就是这样。

我们写代码，真的不要单纯的为了生计，单纯的觉得这是一个青春饭，我们可以把它当做一个一辈子的事业，30岁以后你转型产品，转型架构师，你都是要有code的积淀的，不是说能转就能转的。

一个一生的事业，我想是值得你付出时间去学习的，鸡汤就这么多了。

对了，电话面试的那个研究生小哥还记得么，他去了阿里，也恭喜他，也恭喜这个老兵读者，我实在没想到我的很多文章确实能帮助到大家这么多，我会继续写的。

我是敖丙，一个在互联网苟且偷生的工具人。

最好的关系是互相成就，各位的**「三连」**就是丙丙创作的最大动力，我们下期见！

> 文章持续更新，可以微信搜索「 **三太子敖丙** 」第一时间阅读，回复【**资料**】【**面试**】【**简历**】有我准备的一线大厂面试资料和简历模板，本文 **GitHub** https://github.com/JavaFamily 已经收录，有大厂面试完整考点，欢迎Star。