个人心得：

在此声明一下，这个纯属个人的一些体会，有不当之处，还望批评指正。我本来就是做 Java 相关工作的。在我 10 来家公司的面试过程当中，大概有这样一个比例：按照 100 分制来划分的话，Java 面试相关的，占到 30-40，大数据部分 60-70。

面试过程中最好将自己知道的用图的形式表现出来，需要写的就写出来。

Java 部分

1、JDK 源码的了解情况，concurrent 包下了解多少，hashMap、concurrentHashMap。

2、多线程。join、wait、notify、notifyAll 等方法的具体功能？手写一个程序实现：两个线程同时打印，A 线程打印 1,3,5,7,9，B 线程打印 2，4,6，8,10,要求显示的结果是有序的，1,2,3,4,5,6,7,8,9,10。至少有三种方式实现。

3、并发程序如何编写？这种问题，问的太空，太大，个人觉得回答这种问题，可以举一个实际的场景。比如电商行业中的秒杀，这种场景怎么做的。然后讲讲 synchronized、lock 之类的对比区别，底层实现机制，我觉得就 OK 了。

4、你的项目是如何架构的？出了问题怎么解决？这个问题可能有的人遇不到，大概的意思就是要有整体意识，不想当将军的士兵不是一个合格的士兵吧，大概就这个意思。

5、谈谈你对 Java 封装、继承、多态的认识？这是一个贝尔实验室的老先生问的，这个问题确实让我不好回答，大家各显神通吧

6、内存溢出和内存泄漏的区别？你在工作中遇到过什么情况，怎么解决的？

7、你对设计模式有了解吗？能写出来吗？

大数据部分：

1、描述下项目架构，各个模块的功能？

2、你对 hadoop 了解多少？能否描述下 HDFS 的读写流程，yarn 的资源调度，看过源码吗？MapReduce 的执行流程？

3、flume 集群搭建，flume 如何保证数据不丢失？

4、Hbase 的 rowkey 如何设计？Hbase 的读写原理？

5、讲一下 spark RDD，spark 应用程序是如何启动的？spark 有几种运行方式？各有什么优缺点？你们使用的是哪一种？

6、spark 算子有哪些？你经常使用的有哪些？

7、简述 RDD、dataframe、dataset 的区别和联系、内部实现、相互转换？

8、spark on yarn 经常出现的问题有哪些？如何处理

9、对海量数据的处理问题

10、spark 应用程序中 executor 分配的数量，集群配置，应用运行时间？

11、hadoop 应用程序优化，spark 应用优化

12、spark shuffle 原理？

13、hadoop WordCount 应用程序的整个执行流程？spark WordCount 应用执行流程分析

14、spark SQL 加载数据源

15、比如网页上有一个咨询功能，现在需要实时统计用户的咨询时长？怎么设计，如何实现？

16、加分项：讲解 spark 算子的实现原理，stage 划分算法，资源调度算法等

17、你们项目的数据量多少，常用统计指标有多少个？

18、kafka 如何保证数据有序？如何保证数据不丢失？有多少个 partition？

19、知道闰秒吗？

20、spark 应用内存溢出发生在什么情况下?哪些算子会产生内存溢出？如何定位？如何解决？

21、谈谈对分布式的理解？

22、对机器学习的了解，KNN 算法的实现和应用场景？基本上都是这个套路，工作中有没有使用过？