1. 乐视

第一次面试，准备不足。

技术性问题：

1. Hadoop体系
2. MR的过程
3. SparkSQL和Spark架构，运行流程图，Spark运行的两种方式。常用的Spark函数有哪些？GroupByKey的作用？
4. HIVE中UDF UDAF UDTF的区别。数据倾斜问题怎么解决？表连接有几种类型？
5. HIVE怎么优化？HIVE常用的几个配置是哪些？
6. Shell编程获取时间，crontab五个星号代表什么？Sed和Awk程序的编写
7. TalkingData

技术性问题：Spark了解多少？

非技术性：是否愿意出差，平时爱好。。。

1. 网德天下

MapReduce过程，Zookeeper实现高可用的原理，线程池怎么实现

前三家都比较紧张，很被动，准备的也不好。

1. 东方国信

谈的都比较顺利，主要问了项目，之后表示薪资有点高（13k），我没想的再降低，就没有通知复试了。

1. 畅捷通

笔试全是算法和数据结构：二叉查找树打印、求两个集合交集、算法优化、并发和异步编程、设计两个算法。

1. 太极

谈的很深很细。

按简历一行一行跟我谈了差不多两个多小时。问我Zookeeper实现分布式锁用哪个jar包，以及写mr、spark作业程序具体应该用哪些包？

公司大数据用的集群环境非常好，但最后薪资太低没有联系。

1. 北京四中网校

谈项目，之后问了我一些可能是他们遇到的问题，让我写算法解决，没做出来。

1. SenseTime

主要问Flume如何使用。

Java内存错误中的一个错误，问我如何解决。

Callable和Runable的区别

是否参加过CodeReview？有什么心得？

在线项目的bug如何调试？

1. 易观智库

Hive什么时候不能用combiner？

添加自定义UDF的时候都是临时的，怎么永久添加？

Hive动态分区

Shell获取某行第几个字符怎么做

1. 银创科技

Hadoop用的什么版本？他们公司用的是商业版

PV UV IP解释，数据倾斜问题怎么解决，进程和线程的区别，线程如何配置？

1. 京东

问了项目，让我写shell脚本求一个文件中的平均值

1. 壕鑫互连

就问我会不会写PLSQL，好像只是在招写SQL的人。

1. 滴答拼车

问我Hive怎么解决数据倾斜

MapReduce过程

一个100G文件，里面每一个行都是一个ip，怎么写单机程序找出出现次数最多的前100个IP？

写个二分查找

写程序找出两个字符串中出现的共同字母

写一个表的查询的sql语句，具体忘了，是个嵌套的SQL

问题种类：

1. 计算框架类
   1. MR流程
   2. Spark框架
   3. Storm框架
   4. HIVE
2. Shell脚本类
   1. Awk
   2. Sed
   3. Crontab
3. SQL类
4. 算法设计类
   1. 大数据类
      1. 上百G文件的处理和计算
   2. 常用算法
      1. 快速排序
      2. 二分查找
      3. 冒泡排序
5. 概念知识类
   1. Java相关
      1. 线程、进程 Callable Runnable
      2. Java内存模型
      3. JVM优化
      4. Java提供的集合框架相关
   2. 业务相关
      1. PV UV IP
      2. 使用的工具的版本