第一家 喜马拉雅FM

电话面试，问了40多分钟

第二家 中星电子

没有笔试题，直接就问项目，问了40多分钟

1，两个项目电信和交通厅，分别用了什么架构，怎么搞得，参与搭建了吗？

2，接着又问flume几台，怎么从其他系统获取的数据，kafka几台？

3，我说的kafka吞吐量10万条信息每秒，我们用了一台，接着问那一台kafka挂了呢？

4，这个地方回答的不好，没搞过kafka高可用，说多台kafka也是坑，到处都是陷阱。

5，项目中那块是你做的？我说的storm实时通话分析那里，问storm怎么从kafka里读数据的，

6，接着又问storm的spout几台，我说一个，接着说spout挂了怎么办？实在没法回答这些破问题，根本都没遇到过，吹的话那继续深入的问，一堆坑。

7，storm处理完数据写入哪里？回答hdfs和hbase 又问storm怎么写入hdfs和hbase的具体说一下。

8，看我写的有私有云，问了下openstack，我回答了下openstack的各个模块和作用

9，问我交通厅项目主要做了哪些部分？我说spark MLlib预测路况那部分，问用的什么算法，我说逻辑线性回归

10，接着问线性回归的原理，什么场景适合线性回归，举两个例子说下。

11，模型生成完以后是怎么知道预测的好坏的？

12，对了还问了storm处理的时候利用率怎么样？怎么检测storm没有问题的，程序跑通就一定没有问题吗？反正我也不知道怎么回答了，不知道大数据有没有测试人员，怎么测试改需求。

第三家 银橙传媒

没有笔试，直接问项目，问了一个多小时，问得我他妈的都要吐了

1，自我介绍，然后就项目，电信项目我主要做了那一块？我说strom实时通话分析那块，

2，怎么从其他系统获取的数据，回答flume+kafka+storm这样的流程。

3，接着问flume有几台，通过什么协议获取的数据，然后就开始开火了，

4，flume收集信息的时候遇到了什么问题？怎么解决的？

5，kafka几台，我回答一台，因为kafka最大支持吞吐量10万条每秒，接着问你们kafka传输的实际吞吐量是多少条每秒，一直追问这个，我没遇到过真不知道怎么回答，kafka传输数据的时候遇到什么错误吗？怎么解决的？又是坑，说没有遇到过。接着又问你们kafka处理的时候都没遇到过什么问题吗？弄得我无言以对，沉默。

6，kafka为什么可以支持那么大的吞吐量，怎么实现的，我直接说不知道。

7，看我写hbase+solr，让我说了一下，解释了一下，问solr你们怎么建索引的，建了多少索引，根据什么去建的这些索引？最后还问solr的索引表是单张表，还是多张表，这些表是存在hbase里面了还是分开的？为什么没有存hbase里面，搞得我无言以对。

8，storm和spark streaming的区别？

9，spark MLlib那部分也问了我很多，因为他没搞过机器学习，所以这部分回答的问题不大。

10，出了个题让我现场搞一下，一个文件里有两个字段分别是IP和time，ip可以通过写好的函数转换成省份，让我实时统计每一分钟的PV，UV，注意PV,UV是累加的，不是一分钟的数据，可以根据省份去查询PV，UV。

11，看我写掌握redis，直接问我，redis支持的最大数据量是多少？redis集群下怎么从某一台集群查key-value。

12，问linux里一个文件，怎么替换某个单词的内容，一个文件有几行数据，怎么直接查看第二行数据。

<http://www.talkingdata.com/game/document/zh_cn/index.jsp#Overview>

