7.30 一面：
1.自我介绍，聊了一会儿项目
2.介绍一下 MR 的 shuffle 机制
3.编程题:
有一文件 a.txt 文件内容为每行由字符串数字用逗号分隔，
例:
abc,1
abc,3
lsl,9
ash,10
flu,11
…
现在要求实现一个单机程序，统计出这个文件中相同字符串出现的次数和对应数字的总和

用 java 统计词频问题，，我用了 HashMap 来统计
4.场景设计，还是刚才那道题，如果有文件中一亿条数据，怎么办？
我回答考虑多线程，还是用 Map，但是不能用 HashMap，需要用多个 ConcurrentHashMap，使用分区锁，能够保证线程安全，将线程分为 Map 线程和 Reduce 线程，Map 线程用来 map 数据，
reduce 线程用来进行数据合并。
5.多个线程怎么读取文件，都是从头读吗？
可以用 bufferedReader 按照字节分开读。
6.那你怎么保证读整行的数据呢？
可以通过\n’来判断是不是一行，如果不是知道找到前面是‘\n’的字节
5.reduce 任务什么时候开始呢？
只要有 map 任务完成，就可以开始 reduce 任务
6.用过 Spark 吗？刚才的词频统计用 spark 怎么实现？
用 scala 写的，就一行代码：
line.map(_.split(",")).map(x=>(x(0),x(1).toint,1)).reduceBykey((x,y)=>(x._1+y._1,x._2+y._2))

8.8 二面：
1.自我介绍，其中一个项目是一个科研项目，面试官估计比较感兴趣，唠了一会儿
2. MR 提交 job 到 YARN 的流程
3.MR 运行过程中会发生 OOM，OOM 发生的位置？
4.比如 Hive 任务报 OOM，如何进行优化?
我回答了对 Map Task 数量进行重新配置，以及影响 Map Task 数量的几个参数，块大小，切片大小，默认 mapper 数量等。
5.YARN 的有哪些调度策略？讲一讲 Fair 调度策略？哪一种调度策略会发生饿死？capacity
FIFO 、Fair、Capacity
6.编程题：
给定一个字符串
判断是否为 IP 格式
是=>True

不是=>False
用 python 写正则表达式，一行搞定，无奈忘记 ip 地址的正则表达式怎么写了
就硬着头皮来，先按照”.“切成段，不等于四段的话肯定就不是了
是四段的话再把每一段转成数字判断是否在 0-255 之间，当时有点投机了，我直接用 parseInt 转数字了，不能转我抛出个异常，这点让面试官不太满意
不过后来我又解释了可以用 ASCII 码判断是不是数字并把串转成数字
7.问了我写的算法的复杂度
8.问了点 kafka 的问题，遇到过什么问题？
9.场景：使用 kafka 时候发生宕机，重启后怎么从上次消费的地方接着处理？
不太清楚，但是我觉得 kafka 是高可用容灾备份机制的
10.解释一下 mysql 的索引？索引为什么会快？
11.如何判断查询是否命中索引？