杉数科技：（笔试+面试两小时）

笔试题：

1、数据量未知的文件，如何实现对其进行排序。

2、两个文件，50亿URL，每条50K左右(记不清了)，内存4G，找出两个文件中共有的点击次数前五的URL。

3、给出文件格式如下：

2018,12,23,10

2018,12,24,-10

2012,12,1,30

2012,12,15,20

要求输出文件格式为(需要实现二次排序)

2012-12 30,20

2018-12 10,-10

面试题：

1、简单介绍一下项目

2、你们项目有哪些模块

3、你们spark项目是使用spark-on-yarn还是standalone方式

4、你们提交运行的脚本如何编写的（我答有人写shell脚本，不建议之后这样答）

5、如果提交的任务出现bug，而提交的人不是你，你们怎么进行沟通协调

6、简单说明一下MR的shuffle过程

7、给定两张表，手写MR进行join操作

表一：NAME，SEX 表二：NAME，AGE

8、还是上面的表，手写Spark进行join操作，要求不使用join算子，spark的shuffle和MR的shuffle有什么区别

9、上面你在进行spark的join操作时用到了map算子，我可不可以用foreach算子替代

10、action类算子和tranform类算子哪一种更容易引起OOM，Spark算子调优问题

11、手写链表的增删改查

12、单例设计模式简述，并说明适用情景

13、你在开发Spark的时候调用的tuple能解释下么，或者说我们为什么用tuple不用map，这两者有什么区别

14、python了解多少

15、有没有集群的搭建经验（实机的），有没有租用云平台搭建的经验

16、你对我们公司有多少了解，为何选择我们公司，有什么未来的规划