

1.在使用hadoop或者是spark中遇到过哪些问题，是如何处理解决的？

(1)数据倾斜。

出现这种情况：多数是由于代码的质量写的不够健壮。查看日志：发现问题。

(2) spark-出现OOM

小数据量的情况可以cache，数据量大的情况必须考虑内存使用。

2.公司里集群规模。hbase数据量。数据规模。

3.看过源码？你熟悉哪几个部分的源码？

4.对于Spark你觉得他对于现有大数据的现状的优势和劣势在哪里？