1.hbase与mysql的区别

2.spark实时作业宕掉，kafka指定的topic数据堆积怎么办？

3.java的设计模式知道么，项目中运用在哪？

4.开发中遇到的最困难的事，你是怎么解决的？

5.集群规模？集群配置？

6.spark作业有多少个？

7.项目中当前batch获取前面的batch的算子是哪个？

8.hive开窗函数写过没有？怎么写的？除了row\_number还知道哪些窗口函数？（就那四个排序窗口函数）他们有啥区别？

9.100个分片，我想聚合成两个分片，用哪个算子？-（当时忘了，但是记得是C开头的一个算子，我就说我不会读那个单词，是c开头的算子）--coalesce

10.spark的作业提交流程

11.公司原来大数据部门规模

12.flatmap和map的区别

13.spark对接hbase

14，dateframe和dataset区别