1.介绍下你在公司的项目，你负责哪块？

2.你擅长哪些技术，项目中遇到过哪些问题？

3.spark-on-yarn任务启动，任务提交，资源分配划个流程图

4.iava代码、sql语句类问题

5.cdh用的什么版本，组件熟悉哪些，怎么升级cdh

6.spark从kafka拉取数据时，offset变量保存在哪里？

7.spark运行的三种模式是什么，你们公司在用那种，client和cluster有什么区别，你们公司用的哪个？

8.详细MR和spark有什么区别？

9.为什么说rdd是弹性分布式数据集？

10.hive数据仓库理解吗，简单说一下数据仓库的搭建

11.kafka拉取的超时数据如何处理的。

12.什么是shuffle，如何处理数据倾斜？

13.硬件资源一定的情况下，如何处理海量数据，保证执行速度?

14.如何判断一个循环链表？

15.你们公司大数据部门多少人，你们和那个部门对接，系统是谁在用

16.数据结构，算法方向了解嘛，简单的说一下。