大数据面试题

1. mapreduce 过程
2. hbase 和传统数据库的区别
3. hbase 读数据过程
4. hbase master 和 regionserver 的交互
5. hbase 的 ha，zookeeper 在其中的作用，master 宕机的时候，哪些能正常工作，读写数据？region 分裂？
6. 数据倾斜
7. mysql 索引，哪些索引？实现原理？哪些存储引擎支持 B 树索引，哪些支持 hash 索引？
8. 为啥 mysql 索引要用 B+树而 MongoDB 用 B 树？
9. Mysql 查询优化？
10. 主键和唯一索引的区别
11. 事务的隔离机制，mysql 默认是哪一级
12. MyISAM 和 InnoDB 存储引擎的区别

13. mysql 查询优化，慢查询怎么去定位？

14. mysql 中的各种锁，乐观锁，悲观锁（排他锁，共享锁）；行锁，表锁是怎么实现的？

15. mapreduce 支持哪些 join，map 端？reduce 端？semi join？semi join 你可以通过什么算法去优化？

16. mapreduce 实现二次排序

17. 用 mapreduce 实现两表 join

18. 用 mapreduce 实现一个存储 kv 数据的文件，对里面的 v 进行全量排序

19. zookeeper 实现原理，zab 协议以及原子广播协议

20. paxos 协议，multi-paxos，zab，raft 各种分布式协议内容，使用场景

21. hadoop namenode 的 ha，主备切换实现原理，日志同步原理，QJM 中用到的分布式一致性算法（就是 paxos 算法）

22. spark 运行架构

23. spark 运行原理，从提交一个 jar 到最后返回结果，整个过程

24. spark 的 stage 划分是怎么实现的？拓扑排序？怎么实现？还有什么算法实现？

25. spark rpc，spark2.0 为啥舍弃了 akka，而用 netty?

26. spark 的各种 shuffle，与 mapreduce 的对比

27. spark 的各种 ha，master 的 ha，worker 的 ha，executor 的 ha，driver 的 ha,task 的 ha,在容错的时候对集群或是 task 有什么影响？

28. spark 的内存管理机制，spark1.6 前后对比分析

29. spark2.0 做出了哪些优化？tungsten 引擎？cpu 与内存两个方面分别说明

30. spark rdd、dataframe、dataset 区别

31. callable runnable 区别

32. synchronized 与 lock 区别

33. 类加载机制

34. gc 算法

35. spark 数据倾斜

36. spark shuffle

37. spark 内存管理

38. 各种排序算法，时间复杂度，空间复杂度，spark 和 hadoop 中 shuffle 中各个阶段用到的排序算法把这几种排序算法的使用场景。