**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。