Hadoop面试题系列(9/11)

一. 问答题

1. 用mapreduce实现sql语句select count(x) from a group by b ?

2. 简述MapReduce大致流程，map -> shuffle -> reduce

3. HDFS如何定位replica

4. Hadoop参数调优: cluster level: JVM, map/reduce slots, job level: reducer, memory, use combiner? use compression?

5. hadoop运行的原理 ?

6. mapreduce的原理 ?

7. HDFS存储的机制 ?

8. 如何确认Hadoop集群的健康状况 ?

二. 思考题

现有1 亿个整数均匀分布，如果要得到前1K 个最大的数，求最优的算法。（先不考虑内存的限制，也不考虑读写外存，时间复杂度最少的算法即为最优算法）

我先说下我的想法:分块，比如分1W块，每块1W个，然后分别找出每块最大值，从这最大的1W个值中找最大1K个，那么其他的9K 个最大值所在的块即可扔掉，从剩下的最大的1K
个值所在的块中找前1K个即可。那么原问题的规模就缩小到了1/10。
问题：
(1) 这种分块方法的最优时间复杂度。
(2) 如何分块达到最优。比如也可分10W 块，每块1000
个数。则问题规模可降到原来1/100。但事实上复杂度并没降低。
(3) 还有没更好更优的方法解决这个问题。