# 海量数据

算法思想：分而治之+Hash

1. IP地址最多有2^32=4G种取值情况，所以不能完全加载到内存中处理；   
   2.可以考虑采用“分而治之”的思想，按照IP地址的Hash(IP)%1024值，把海量IP日志分别存储到1024个小文件中。这样，每个小文件最多包含4MB个IP地址；   
   3.对于每一个小文件，可以构建一个IP为key，出现次数为value的Hash map，同时记录当前出现次数最多的那个IP地址；  
   4.可以得到1024个小文件中的出现次数最多的IP，再依据常规的排序算法得到总体上出现次数最多的IP；

cat 222.txt | awk -F '//' '{print $3}' | sort | uniq -c | sort -k1nr | head -3

获取访问量最大的3个url

Topk 这种算法是用堆, 时间复杂度是nlogk,

为啥不是klogn呢

是这样的,我们先建一个大小为k的堆,然后n个数据都往里面去比较,首先,如果你的数组大小就是k了,那么每次调整的次数就是logk咯.

如果量很大,超过内存了,反正就是把这些内容先遍历,然后做hash算法分到100000个小文件里,然后操作这些小文件,这不就是mapreduce,分治的想法么.