1. 海量日志数据，提取某日访问百度次数最多的那个IP？

分而治之+hash：

①IP地址最多有2^32个，总大小为4G

②按照IP地址的HASH（ip）映射到1024个小文件中，每个文件4M

③对于小文件可以找到频率最高的IP

④得到1024个出现较多的IP，再排序一下得到访问最多的IP

1. 有一个1G大小的文件，里面每一行是一个词，词的大小不超过16字节，内存限制是1M。请返回频率最高的100个词？

顺序读文件，对每个词去hash(x)%5000，得到5000个小文件。每个文件大概200k，如果文件中有超过1M的，可以继续hash下去分直到所有小文件都<1M。

对于每个小文件统计词&词频（trie树或者hash表），并取出频率最大的100个词（可以用最小堆），并把100个词存入文件，这样就又得到5000个文件，下一步就是对5000个文件进行归并（类似归并排序）。

1. 有10个文件，每个文件1G，每个文件的每一行存放是都是用户的query，每个文件的query都有可能重复。要求按照query的频率排序。

方案①顺序读取10个文件，按照hash(query)将结果写入10个文件中，大约1G，找一台2G内存的服务器对其进行query频率统计并排序，重新输入到新的10个文件中，最后对这10个文件进行归并排序（内排序+外排序）。

方案②一般来说query的总量是有限的，只是重复的很多而已，对于所有的query，可以一次性加载到内存，这样就可以用hash表或者trie树来直接统计query频率，最后按照频率来排序就好了。

方案③mapreduce

1. 在2.5亿个整数中找出不重复的整数，注意内存不足以容纳2.5亿个整数？

方案①位图2-Bitmap（每个数分配2bit，00表示不存在，01表示出现1次，10表示多次，11表示无意义），共需要2^32\*2bit=1G内存。