**1、海量日志数据，提取出某日访问百度次数最多的那个IP。【1.分而治之/hash映射 + hash统计 +堆/快速/归并排序；】**

**2.搜索引擎会通过日志文件把用户每次检索使用的所有检索串都记录下来，每个查询串的长度为1-255字节。**假设目前有一千万个记录（这些查询串的重复度比较高，虽然总数是1千万，但如果除去重复后，不超过3百万个。一个查询串的重复度越高，说明查询它的用户越多，也就是越热门。），请你统计最热门的10个查询串，要求使用的内存不能超过1G。

100\*300\*10000byte = 0.3g **hash统计+堆排（小根堆）/trie树+堆排（小根堆）**

1. **有一个1G大小的一个文件，里面每一行是一个词，词的大小不超过16字节，内存限制大小是1M。返回频数最高的100个词。**

**Hash映射为5000个小文件，每个大约200k.如果还有大于1M的文件就继续hash，直到全部不大于1M。然后将每个文件的词hash统计，再用小根堆选出每个文件的top100，并按频率排序，然后写进5000个文件（8\*1000\*1000字节××）。最后文件之间进行2路归并排序，直到归并为1个大文件为止。**

1. **海量数据分布在100台电脑中，想个办法高效统计出这批数据的TOP10。**
2. **有10个文件，每个文件1G，每个文件的每一行存放的都是用户的query，每个文件的query都可能重复。要求你按照query的频度排序。**

**②哈希统计+归并/堆/快速排序（虽然文件大，但是query重复较多，可能可以一次加入内存）**

1. **给定a、b两个文件，各存放50亿个url，每个url各占64字节，内存限制是4G，让你找出a、b文件共同的url？**

**5G\*64=320G，所以将a、b分别映射为1000个小文件，大约300M。这样相同的URL都在对应文件中。所以求出1000对文件中的相同URL。即将文件1数据放进set，遍历对应文件2看数据是否在set里，在就写入最终一个新文件。。**

**7、怎么在海量数据中找出重复次数最多的一个？**

**哈希映射+哈希统计+归并/快速/堆排序。（类似第1题）**

1. **上千万或上亿数据（有重复），统计其中出现次数最多的N个数据。**

**哈希映射+哈希统计+小根堆+排序+2路归并+小根堆**

**9、一个文本文件，大约有一万行，每行一个词，要求统计出其中最频繁出现的前10个词，请给出思想，给出时间复杂度分析。**

**Trie树+小根堆 max（O(n\*le)，O(n \*log2k）**

1. **1000万字符串，其中有些是重复的，需要把重复的全部去掉，保留没有重复的字符串。请怎么设计和实现？**

**①先哈希映射成1000个文件，依次将每个文件字符串插入到set中，再写入指定的新文件。**

**②先哈希映射成1000个文件，依次将每个文件字符串hashmap统计，再写入指定文件。**

**11、2.5亿个整数中找出不重复的整数的个数，内存空间不足以容纳这2.5亿个整数。**

**先哈希成几个文件，每个文件采用2bitmap（00是0次，01是1次，10是多次，11无效）。**

**然后遍历字符串，00变01,01变10,10不变。 最后遍历bitmap将对应位是01的数输出。**

**12、5亿个int找它们的中位数。**

**13、腾讯面试题：给40亿个不重复的unsigned int的整数，没排过序的，然后再给一个数，如何快速判断这个数是否在那40亿个数当中？**

**①快排+二分查找**

**②申请512M内存，1bit表示一个unsigned int.读入数据，将数据相应的位置1.**

**③bloom fitter**

**④先将大文件分为最高位1和0的文件，大小为<=20亿和>20亿。与此数最高位比较进入相应文件再查找，分为次高位1和0的文件，大小为<=10亿和>10亿。与此数最高位比较进入相应文件再查找。以此类推，o(log2n)**

**14、题目：给你100000个长度不超过10的单词。对于每一个单词，我们要判断他出没出现过，如果出现了，求第一次出现在第几个位置。**

**Tire树（以空间换时间，利用字符串公共前缀节省查询时间）**

**15、在一个二维数组中，每一行都按照从左到右递增的顺序排序，每一列都按照从上到下递增的顺序排序。请完成一个函数，输入这样的一个二维数组和一个整数，判断数组中是否含有该整数。**  
例如下面的二维数组就是每行、每列都递增排序。如果在这个数组中查找数字6，则返回true；如果查找数字5，由于数组不含有该数字，则返回false。

本Young问题解法有二（如查找数字6）：

**16、输入：一个最多含有n个不重复的正整数（也就是说可能含有少于n个不重复正整数）的文件，其中每个数都小于等于n，且n=10^7。  
输出：得到按从小到大升序排列的包含所有输入的整数的列表。  
条件：最多有大约1MB的内存空间可用，但磁盘空间足够。且要求运行时间在5分钟以下，10秒为最佳结果。**

**1、先内存排序**由于要求的可用内存为1MB，那么每次可以在内存中对250K的数据进行排序，然后将有序的数写入硬盘。  
那么10M的数据需要循环40次，最终产生40个有序的文件。  
**2、后归并排序**

1. 将每个文件最开始的数读入(由于有序，所以为该文件最小数)，存放在一个大小为40的first\_data数组中；
2. 选择first\_data数组中最小的数min\_data，及其对应的文件索引index；
3. 将first\_data数组中最小的数写入文件result，然后更新数组first\_data(根据index读取该文件下一个数代替原来min\_data的位置)；
4. 判断是否所有数据都读取完毕，没有就返回2。

**10. 100w个数中找出最大的100个数。**

**方案1：**

在前面的题中，我们已经提到了，用一个含**100个元素的最小堆**完成。复杂度为O(100w\*lg100)。