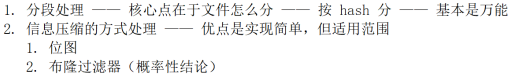
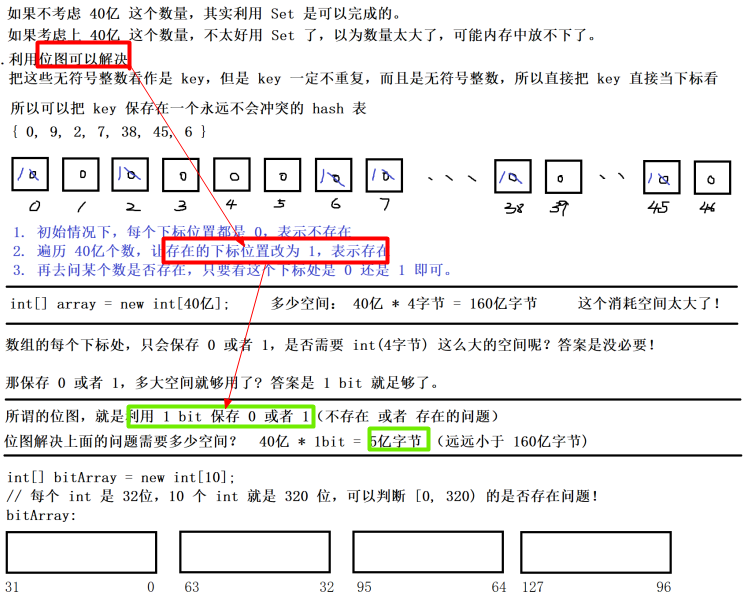
海量数据处理：

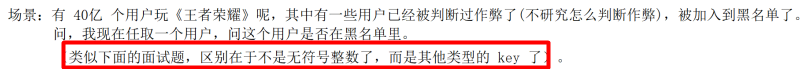


**1.给定40亿个不重复的无符号数，没排过序。给一个无符号数，如何快速判断一个数是否在这40亿个数中？**

**第一种:位图**

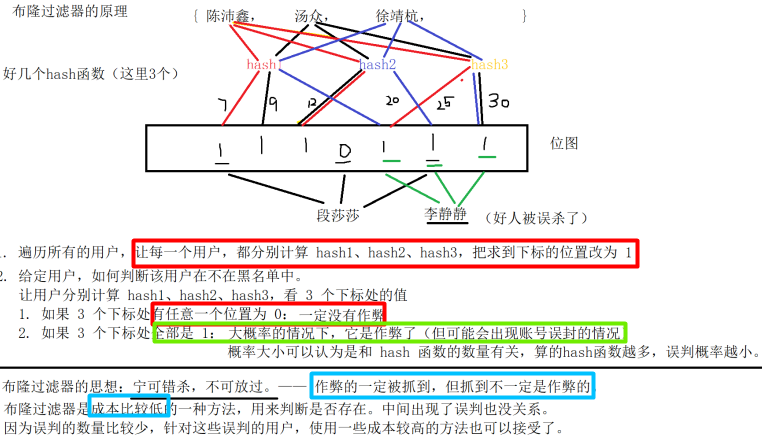


**第二种：布隆过滤器**



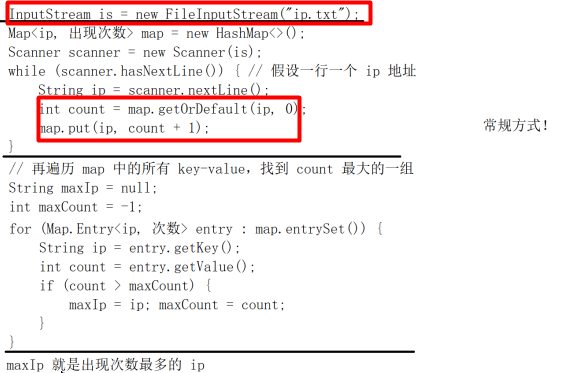
clipboard.png

**原理：**



**2.给一个超过100G大小的log file,log中存着IP地址，设计算法找到出现次数最多的ip地址？**

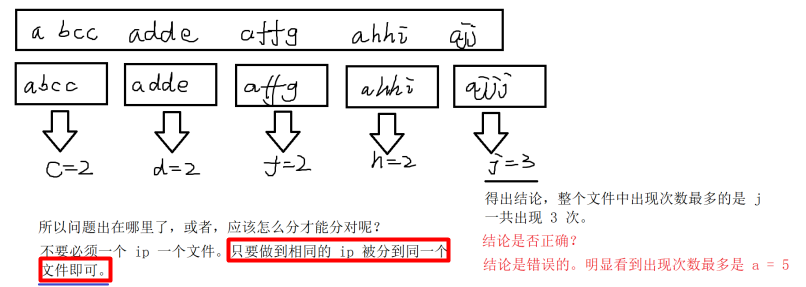
**不考虑内存不够的情况下：采用Map进行解决，找到出现次数最多的ip**



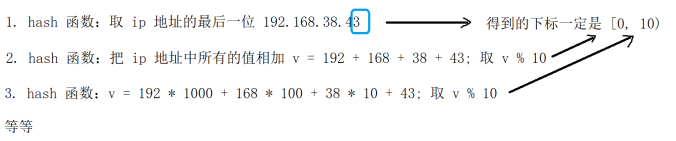
**考虑内存不够的情况：将文件进行拆分**

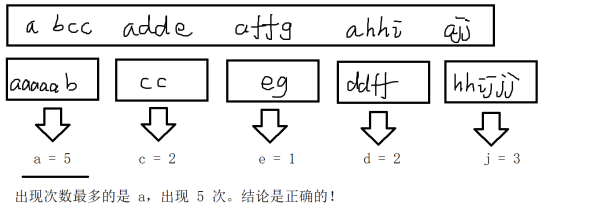
如何拆分？

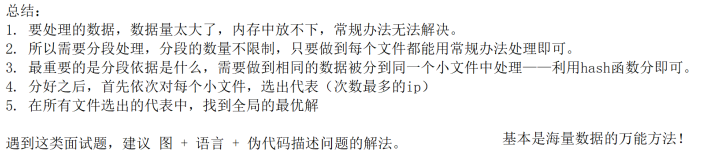
**随便分：**



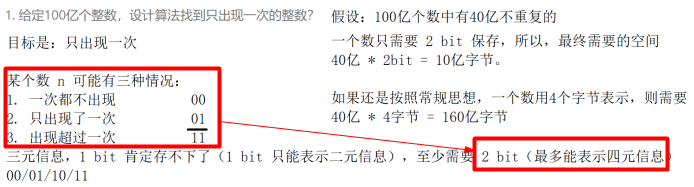
**需要使得相同ip被分到同一个文件（相同的ip==相同的分组编号Hash函数）**







**3.给定100亿整数，设计算法找到只出现一次的整数？**



**4.给定两个文件，分别有100亿个整数，只有1G内存，如何找到两个文件交集？**

