**海量数据处理算法：**

**Bloom filter**:用位数组简洁的表示一个集合，并能判断一个元素是否属于这个集合，实现是用k个哈希函数将元素映射到一个m长度的数组。是一个判断元素是否属于集合的算法，可能会误判，但不会漏判

（时间空间互换之外，引入了错误率的概念，以错误率换取时空效率）

**Hash**:hash+链表

**Bit-map**:所谓的[Bit-map](http://blog.csdn.net/hguisu/article/details/7880288" \t "https://www.cnblogs.com/ECJTUACM-873284962/p/_blank)就是用一个bit位来标记某个元素对应的Value， 而Key即是该元素。由于采用了Bit为单位来存储数据，因此在存储空间方面，可以大大节省。可进行数据的快速查找，判重，删除，一般来说数据范围是int的10倍以下

0-7内的5个元素(4,7,2,5,3)排序

开辟1byte的空间，值全部标为0，将第4,7,2,5,3位标为1，顺序读出为1的位置，23456排序完成

**双层桶**：面对一堆大量的数据我们无法处理的时候，我们可以将其分成一个个小的单元，然后根据一定的策略来处理这些小单元，从而达到目的。

【问题实例】   
1).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。

有点像鸽巢原理，整数个数为2^32,也就是，我们可以将这2^32个数，划分为2^8=256个区域(比如用单个文件代表一个区域)，然后将数据分离到不同的区 域，然后不同的区域在利用bitmap就可以直接解决了。也就是说只要有足够的磁盘空间，就可以很方便的解决。

2).5亿个int找它们的中位数。

这个例子比上面那个更明显。首先我们将int划分为2^16个区域，然后读取数据统计落到各个区域里的数的个数，之后我们根据统计结果就可以判断中位数落到那个区域，同时知道这个区域中的第几大数刚好是中位数。然后第二次扫描我们只统计落在这个区域中的那些数就可以了。

**倒排索引**：VSM全称是Vector Space Model(向量空间模型)，是IR(Information Retrieval信息检索)模型中的一种，由于其简单，直观，高效，所以被广泛的应用到搜索引擎的[架构](http://lib.csdn.net/base/architecture" \o "大型网站架构知识库" \t "https://www.cnblogs.com/ECJTUACM-873284962/p/_blank)中。98年的Google就是凭借这样的一个模型，开始了它的疯狂扩张之路。

文档(Document)：一个完整的信息单元，对应的搜索引擎系统里，就是指一个个的网页。

标引项(Term)：文档的基本构成单位，例如在英文中可以看做是一个单词，在中文中可以看作一个词语。

查询(Query)：一个用户的输入，一般由多个Term构成。

那么用一句话概况搜索引擎所做的事情就是：对于用户输入的Query，找到最相似的Document返回给用户。而这正是IR模型所解决的问题：

信息检索模型是指如何对查询和文档进行表示，然后对它们进行相似度计算的框架和方法。

相似算法由cos夹角法演变，并结合tf,idf