# 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？

方案 1：可以估计每个文件安的大小为 50G×64=320G，远远大于内存限制的 4G。所以不可能将其完全加载到内存中处理。考虑采取分而治之的方法。

s 遍历文件 a，对每个 url 求取 ，然后根据所取得的值将 url 分 别存储到 1000 个小文件（记为 ）中。这样每个小文件的大约为 300M。

s 遍历文件 b，采取和 a 相同的方式将 url 分别存储到 1000 各小文件（记为

）。这样处理后，所有可能相同的 url 都在对应的小文件

（ ）中，不对应的小文件不可能有相同的 url。然后 我们只要求出 1000 对小文件中相同的 url 即可。

s 求每对小文件中相同的 url 时，可以把其中一个小文件的 url 存储到 hash\_set 中。然后遍历另一个小文件的每个 url，看其是否在刚才构建的 hash\_set 中，如果是，那么就是共同的 url，存到文件里面就可以了。

方案 2：如果允许有一定的错误率，可以使用 Bloom filter，4G 内存大概可以 表示 340 亿 bit。将其中一个文件中的 url 使用 Bloom filter 映射为这 340 亿

bit，然后挨个读取另外一个文件的 url，检查是否与 Bloom filter，如果是，那 么该 url 应该是共同的 url（注意会有一定的错误率）。