大厂面试中经常漫聊的有趣算法问题

前置知识
讲解025、讲解026、讲解027 - 堆结构
讲解031、讲解032 - 位运算、位图
讲解016 - 哈希函数

本节课讲述：
蓄水池采样（有代码）
内存限制类问题（纯聊）
文件排序问题（纯聊）
热词问题（纯聊）
多线程任务分配问题（纯聊）
跳出思维定式的系统设计（纯聊）
囚徒生存问题（有代码）

左程云

大厂面试中经常漫聊的有趣算法问题

题目1
蓄水池采样
假设有一个不停吐出球的机器，每次吐出1号球、2号球、3号球...
有一个袋子只能装下10个球，每次机器吐出的球，要么放入袋子，要么永远扔掉
如何做到机器吐出每一个球之后，所有吐出的球都等概率被放进袋子里

扩展：
如何设计一个抽奖系统，一天内所有登录的用户都有均等的中奖机会，一共100人中奖

左程云

大厂面试中经常漫聊的有趣算法问题

题目2
内存限制类问题
32位无符号整数的范围是0 ~ 4,294,967,295。现在有一个正好包含40亿个无符号整数的文件
可以使用最多1GB的内存，怎么找到出现次数最多的数
本题为上节课的内容，讲解106的视频从 16分30秒 ~ 30分30秒

32位无符号整数的范围是0 ~ 4,294,967,295。现在有一个正好包含40亿个无符号整数的文件
1，内存只有1GB，找到所有没出现过的数字
2，内存只有1GB，找出所有出现了两次的数
3，内存只有为3KB，只用找到一个没出现过的数字
4，内存只有几个变量的空间，只用找到一个没出现过的数字
5，内存只有3KB，找到这40亿个整数的上中位数

大厂面试中经常漫聊的有趣算法问题

题目3
文件排序问题
32位无符号整数的范围是0 ~ 4,294,967,295
有一个10G大小的文件
每一行都装着这种类型的数字
整个文件是无序的
给你5G的内存空间
请你输出一个10G大小的文件是原文件所有数字排序的结果

大厂面试中经常漫聊的有趣算法问题

题目4
热词问题
某搜索公司一天的用户搜索词汇是100亿规模的数据量
每个搜索词汇不超过64字节
请设计一种每天从凌晨0点开始重新统计
当天的任何时刻都能快速显示Top100高频词的可行办法

左程云

大厂面试中经常漫聊的有趣算法问题

题目5
多线程任务分配问题
给定一个List<String> list，每个字符串的长度在1 ~ 1000之间，字符串数量20亿
每个字符串类似："hello,world,have,hello,world"
这个字符串中有2个hello，2个world，1个have
请设计一种多线程处理方案
统计list中每个字符串，切分出来的单词数量，并且汇总
最终返回一个HashMap<String, Integer>表示每个字符串在list中一共出现的次数
机器资源允许你最多开100个线程
可以用多线程读取list列表，也可以用多线程写入HashMap
怎么设计一种多线程的处理流程，让完成目标的速度尽量快

大厂面试中经常漫聊的有趣算法问题

题目6
跳出思维定式的系统设计
请设计一个生成uuid的系统，要求理论上决不允许出现相同的id
该系统供全球使用，需要支持每秒数百万亿级别的并发
你的系统只需要源源不断生成uuid即可，除此之外不需要维持任何关系
该系统如何设计？

左程云

大厂面试中经常漫聊的有趣算法问题

题目7
囚徒生存问题
有100个犯人被关在监狱，犯人编号0~99，监狱长准备了100个盒子，盒子编号0~99
这100个盒子排成一排，放在一个房间里面，盒子编号从左往右有序排列
最开始时，每个犯人的编号放在每个盒子里，两种编号一一对应，监狱长构思了一个处决犯人的计划
监狱长打开了很多盒子，并交换了盒子里犯人的编号
交换行为完全随机，但依然保持每个盒子都有一个犯人编号
监狱长规定，每个犯人单独进入房间，可以打开50个盒子，寻找自己的编号
该犯人全程无法和其他犯人进行任何交流，并且不能交换盒子中的编号，只能打开查看
寻找过程结束后把所有盒子关上，走出房间，然后下一个犯人再进入房间，重复上述过程
监狱长规定，每个犯人在尝试50次的过程中，都需要找到自己的编号
如果有任何一个犯人没有做到这一点，100个犯人全部处决
所有犯人在一起交谈的时机只能发生在游戏开始之前，游戏一旦开始直到最后一个人结束都无法交流
请尽量制定一个让所有犯人存活概率最大的策略