哈希函数、哈希表、布隆过滤器、一致性哈希

前置知识
讲解105 - 哈希函数特征，讲解105的视频 1分06秒 ~ 13分06秒

本节课讲述：
工程上常用的哈希函数展示
哈希值根据余数分组的均匀性
哈希表原理
布隆过滤器原理
一致性哈希原理

左程云

哈希函数、哈希表、布隆过滤器、一致性哈希

上节课内容，讲解105的视频 1分06秒 ~ 13分06秒

哈希函数的用处
把复杂样本变成数字，以后复杂样本之间的对比，就变成了数字之间的对比

哈希函数的基本性质
1）输入参数的可能性是无限的，输出的值范围相对有限
2）输入同样的样本一定得到同样的输出值，也就是哈希函数没有任何随机机制
3）输入不同的样本也可能得到同样的输出值，此时叫哈希碰撞
4）输入大量不同的样本，得到的大量输出值，会几乎均匀的分布在整个输出域上
哈希函数的种类很多，但都符合上述性质
性质4是最重要的，哈希碰撞理论上无法避免，但是好的哈希函数会让碰撞几率变的很低
可以把性质4直观理解为：
不管有没有规律，也不管多么相似，总之一堆不同的输入，得到的输出结果从分布上看，熵最大！

哈希函数、哈希表、布隆过滤器、一致性哈希

哈希函数的扩展作用
利用哈希函数的均匀性，可以把样本进行均匀分组

工程上常用的哈希函数
SHA3-512
SHA-1
SHA-384
SHA3-384
SHA-224
SHA-512/256
SHA-256
MD2
SHA-512/224
SHA3-256
SHA-512
SHA3-224
MD5

展示一下用法 & 均匀性

左程云

哈希函数、哈希表、布隆过滤器、一致性哈希

哈希函数相关问题
一台机器上硬盘空间很大，但是内存空间很少只有4G
给定100亿个字符串的文件，每行是一个字符串长100字节，统计哪个字符串出现次数最多
不需要代码实现，聊清楚原理即可

很多工程上的问题都是利用哈希函数把大数据量的样本均匀分散到多台机器上 或者 多个小文件里
哈希函数可以保证同一个样本一定会放在一起，还可以保证把不同种类的样本均匀分开

左程云

哈希函数、哈希表、布隆过滤器、一致性哈希

哈希表原理 & 哈希表扩容
课上重点讲解
哈希表增删改查的均摊复杂度是$O(k)$，k是样本平均长度

如下的细节都可以定制：
初始桶空间，一开始准备多少个桶？
扩容阈值条件，链表长度多少时扩容？
扩容因子，一次增加多少桶空间？
哈希函数选择，简单的哈希函数 *or* 复杂的哈希函数？
桶结构的具体实现，简单链表？开放地址？红黑树？

不管定制什么样的细节，但是哈希表的原理是不变的
所有的不同定制也仅仅是优化常数时间，时间复杂度无法再优化

左程云

哈希函数、哈希表、布隆过滤器、一致性哈希

$$(1-e^{-\frac{nk}{m}})^k \qquad \times\frac{m}{n}$$

去重系统)，有100亿个url需要进入黑名单，每个url有100字节
可以判断任何一个url在不在黑名单内，预期失误率万分之一，内存占用不超过30G
以做黑名单、爬虫去重系统、还能做数据定位

原理 + 失误类型，课上重点讲解

假设数据量为n，预期的失误率为p，布隆过滤器大小和每个样本的大小无关
1，根据n和p，算出布隆过滤器一共需要多少个bit位，向上取整为m，注意m是比特数量，m/8才是字节数
2，根据m和n，算出布隆过滤器应该选择多少个哈希函数，向上取整，记为k
3，根据修正公式，算出真实的失误率p_true，注意第三个公式带入时，用扩大之后的m

证明略
网上帖子很多
wiki也很多
有兴趣可以深入研究

左程云

哈希函数、哈希表、布隆过滤器、一致性哈希

一致性哈希原理
1，一种简单的存储结构介绍，弱点是 增加 or 减少机器，数据迁移的代价是全量的
2，选择哈希key的注意点
3，一致性哈希实现的分布式存储结构，哈希域变环、机器进环的设计
4，一致性哈希的虚拟节点技术可以规避数据倾斜、实现负载均衡、实现负载管理
课上重点讲解

左程云