

高频数据结构——哈希表与堆

主讲人 令狐冲
课程版本 v7.0

数据结构可以认为是一个数据存储集合以及定义在这个集合上的若干操作（功能）
他有如下的三种考法：

考法1：问某种数据结构的基本原理，并要求实现

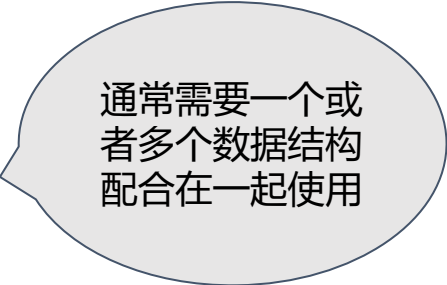
例题：说一下 Hash 的原理并实现一个 Hash 表

考法2：使用某种数据结构完成事情

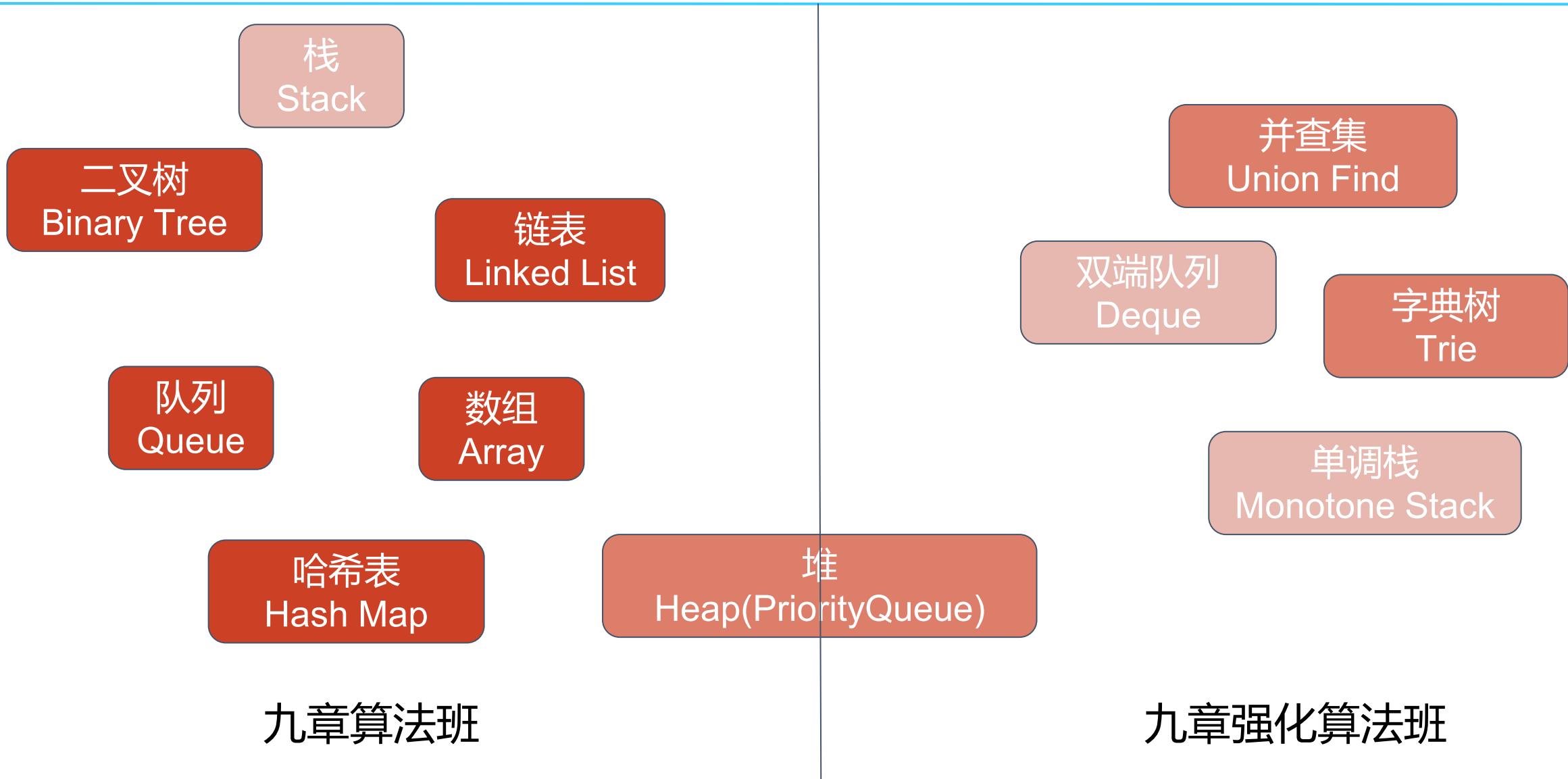
例题：归并 K 个有序数组

考法3：实现一种数据结构，提供一些特定的功能

例题：最高频 K 项问题



通常需要一个或者多个数据结构配合在一起使用



数据结构时间复杂度的衡量方法

数据结构通常会提供“**多个**”对外接口
只用一个时间复杂度是很难对其进行正确评价的
所以通常要对每个接口的时间复杂度进行描述

比如你需要设计一个 Set 的数据结构，提供 lowerBound 和 add 两个方法。lowerBound 的意思是，找到比某个数小的最大值

算法1: $O(n)$ lowerBound $O(1)$ add

使用数组存储，每次打擂台进行比较，插入就直接插入到数组最后面

算法2: $O(\log n)$ lowerBound $O(\log n)$ add

使用红黑树 (Red-black Tree) 存储，Java 里的 TreeSet, C++ 里的 map

上面两个算法谁好谁坏呢？

不一定谁好谁坏！要看这两个方法被调用的频率如何。

如果 lowerBound 很少调用, add 非常频繁, 则算法1好。

如果 lowerBound 和 add 调用的频率都差不多, 或者 lowerBound 被调用得更多, 则算法2好

不过通常来说, 在面试中的问题, 我们会很容易找到一个像算法1这样的实现方法, 其中一个操作时间复杂度很大, 另外一个操作时间复杂度很低。

通常的解决办法都是让快的操作慢一点, 让慢的操作快一点, 这样总体得到一个更快的复杂度。

哈希表 Hash

支持操作: $O(1)$ Insert / $O(1)$ Find / $O(1)$ Delete
问: 这些操作都是 $O(1)$ 的前提条件是什么?

$O(\text{size of key})$

哈希表 (HashMap / unordered_map / dict)

任何操作的时间复杂度从严格意义上来说

都是 $O(\text{size of key})$ 而不是 $O(1)$

你不可能在 $O(1)$ 的时间内判断 2 个 1m 长的字符串是否相等

请在互动课中学习如下先修知识

Hash Table, Hash Map 和 Hash Set 的区别是啥

什么是 Hash Function (产生HashCode的函数), 作用以及实现原理

什么是 Open Hashing 什么是 Closed Hashing

什么是 Rehashing (重哈希)

LRU Cache

<http://www.lintcode.com/problem/lru-cache/>

<http://www.jiuzhang.com/solutions/lru-cache/>

Example: [2 1 3 2 5 3 6 7]

LRU Cache

- Java 中有一个 LinkedHashMap, 本质上是 DoublyLinkedList + HashMap
 - `HashMap<key, DoublyListNode> DoublyListNode {`
 - `prev, next, key, value;`
 - `}`
 - Python 里是 OrderedDict
-
- 新节点从尾部加入
 - 老节点从头部移走

问: Singly List 是否可行?

Singly List 是否可行?

可以, 在 Hash 中存储 Singly List 中的 prev node 即可
如 linked list = dummy->1->2->3->null 时
hash[1] = dummy, hash[2] = node1 ...

Insert Delete GetRandom $O(1)$

<http://www.lintcode.com/problem/insert-delete-getrandom-o1/>

<http://www.jiuzhang.com/solutions/insert-delete-getrandom-o1/>

类似的题: <http://www.lintcode.com/problem/load-balancer/>

* Follow up: 允许重复的数

<http://www.lintcode.com/problem/insert-delete-getrandom-o1-duplicates-allowed/>

<http://www.jiuzhang.com/solutions/insert-delete-getrandom-o1-duplicates-allowed/>

实现较为困难，看懂参考代码即可，不用太纠结

99%的人面试的时候都做不出来，面试时通常给个思路就可以了

Insert Delete GetRandom O(1) 面试评分标准

Strong Hire:

Bug Free 的实现无重复版本的代码，并给出有重复版的基本思路即可

Hire / Weak Hire:

实现无重复版本的代码，无需太多提示或者较少提示，代码 Bug 不多

No Hire / Strong No:

无法无重复版本的给出正确算法，或者无法正确实现，代码 Bug 过多

休息 5 分钟

总结一道题的经验，胜过刷十道题
把你的代码和总结发到九章面试题交流社区
www.jiuzhang.com/solutions

First Unique Number in Data Stream

<http://www.lintcode.com/problem/first-unique-number-in-data-stream/>

<http://www.jiuzhang.com/solutions/first-unique-number-in-data-stream/>

Follow up: 只遍历一次

<http://www.lintcode.com/problem/first-unique-number-in-data-stream-ii/>

<http://www.jiuzhang.com/solutions/first-unique-number-in-data-stream-ii/>

Data Stream 相关问题

<https://www.lintcode.com/problem/?tag=data-stream>

数据流问题 = 数据只能遍历一次

Data Stream 大都和 **Sliding Window** 有关

这类问题我们将在《**九章算法强化班**》中深入讲解

什么是 Data Stream 类问题?

只允许遍历一次!

<http://www.lintcode.com/problem/first-unique-number-in-a-stream-ii/>

<http://www.jiuzhang.com/solutions/first-unique-number-in-a-stream-ii/>

其他 Data Stream 的相关问题:

<http://www.lintcode.com/tag/data-stream/>

哈希表的其他练习题

- <http://www.lintcode.com/problem/subarray-sum/>
- <http://www.lintcode.com/problem/copy-list-with-random-pointer/>
- <http://www.lintcode.com/problem/anagrams/>
- <http://www.lintcode.com/problem/longest-consecutive-sequence/>

Heap

支持操作: $O(\log N)$ Add / $O(\log N)$ Remove / $O(1)$ Min or Max

Heap 的基本原理详见互动课

Java: PriorityQueue

C++: priority_queue

Python: heapq

heapq / PQ vs Heap

主要区别是什么？

heapq / PQ vs Heap

主要区别是什么？

heapq / PQ 的 remove 操作是 $O(n)$ 的

构建一个 heap 的时间复杂度?

是 $O(n)$ 还是 $O(n \log n)$?

构建一个 heap 的时间复杂度?

是 $O(n)$ 还是 $O(n \log n)$?

是 $O(n)$, 用 Heapify

Python: `heapq.heapify(...)`

遍历一个 heap 的时间复杂度?

比如 Java 中可以用 Iterator 来遍历

遍历一个 heap 的时间复杂度?

比如 Java 中可以用 Iterator 来遍历

$O(n \log n)$

Ugly Number II

<http://www.lintcode.com/problem/ugly-number-ii/>

<http://www.jiuzhang.com/solutions/ugly-number-ii/>

在线算法 vs 离线算法

在线算法 = 数据结构设计类问题 = 数据流问题 = 数据不可二次访问 = 多次输入和输出

离线算法 = 一次输入输出 = 数据是一开始给定的 = 数据可以多次访问

Top K 问题离线算法

<http://www.lintcode.com/problem/k-closest-points/>

<http://www.jiuzhang.com/solutions/k-closest-points/>

Microsoft / Apple / Facebook

Top K 问题在线算法

<http://www.lintcode.com/problem/top-k-largest-numbers-ii/>

<http://www.jiuzhang.com/solutions/top-k-largest-number-ii/>

Follow up: Top K Frequent Elements

Top K Largest Number II 面试评分标准

Strong Hire:

能完整实现代码，时间复杂度最优，代码没有大 Bug，无需提示
能够对 Follow up 问题提出解决方案（不一定要实现）

Hire / Weak Hire:

能够完整实现代码，代码 Bug 不多
无需提示或者很少需要提示做出最优复杂度的版本

No Hire / Strong No:

无法用最优的算法实现出来

Related Questions

- <http://www.lintcode.com/en/problem/high-five/>
- <http://www.lintcode.com/problem/merge-k-sorted-arrays/>
- <http://www.lintcode.com/problem/data-stream-median/>
- <http://www.lintcode.com/problem/top-k-largest-numbers/>
- <http://www.lintcode.com/problem/kth-smallest-number-in-sorted-matrix/>

独孤九剑 —— 破掌式

高级数据结构 Cheat Sheet

高级数据结构	各类操作时间复杂度	能够解决哪些问题	考察频率	学习难度	哪里可以学到
Heap 堆	$O(\log n)$ push, pop $O(1)$ top	全局动态找最大找最小	高	低（掌握应用）	九章算法班，强化班
Hash Map 哈希表	$O(1)$ insert, find, delete	查询元素是否存在, key-value 查询问题	高	低（掌握应用）	九章算法班
Trie 前缀树	$O(L)$ insert, find, delete	和哈希表解决问题类似，查询元素是否存在, key-value 查询问题	中	低	九章算法强化班
UnionFind 并查集	$O(1)$ union, find	动态合并集合并判断两个元素是否在同一个集合, MST	中	低	九章算法强化班
Balanced BST 平衡排序二叉树（如红黑树 Red-black Tree）	$O(\log n)$ insert, find, delete, max, min, lower, upper	动态增删查改并支持同时找全局最大最小值 找比某个数大的最小值和比某个数小的最大值时可以用（尽可能接近）	低	高	Google
Skip List 跳跃表	$O(\log n)$ insert, find, delete, max, min, lower, upper	和 Balanced BST 解决的问题一样，并能一直维持一个有序链表	低	高	系统设计班
Binary Indexed Tree 树状数组	$O(\log n)$ insert, delete, range sum	增删改的同时，解决区间求和问题	低	中	树状数组与线段树
Segment Tree 线段树	$O(\log n)$ insert, delete, find, range max, range min, range sum, lower, upper $O(1)$ global max, global min	增删改的同时，解决区间求值问题, max/min/sum 等等 可以完全替代	低	中	树状数组与线段树

- 在第 32 章互动课中继续学习栈和队列的相关面试题
 - 最大栈 / 最小栈
 - 两个队列实现栈 / 两个栈实现队列
- 在第 33 章互动课中继续学习外排序算法与数组合并类问题
 - 外排序算法与 K 路归并算法
 - 数组合并的相关问题