

[阮一峰的网络日志](#) » [首页](#) » [档案](#)分类： [理解计算机](#)上一篇： [Airbnb与创投](#)下一篇： [中文字体网页开发指南](#)

数据库的最简单实现

作者： 阮一峰

所有应用软件之中，数据库可能是最复杂的。

MySQL的手册有3000多页，PostgreSQL的手册有2000多页，Oracle的手册更是比它们相加还要厚。



但是，自己写一个最简单的数据库，做起来并不难。Reddit上面有一个[帖子](#)，只用了几百个字，就把原理讲清楚了。下面是我根据这个帖子整理的内容。

一、数据以文本形式保存

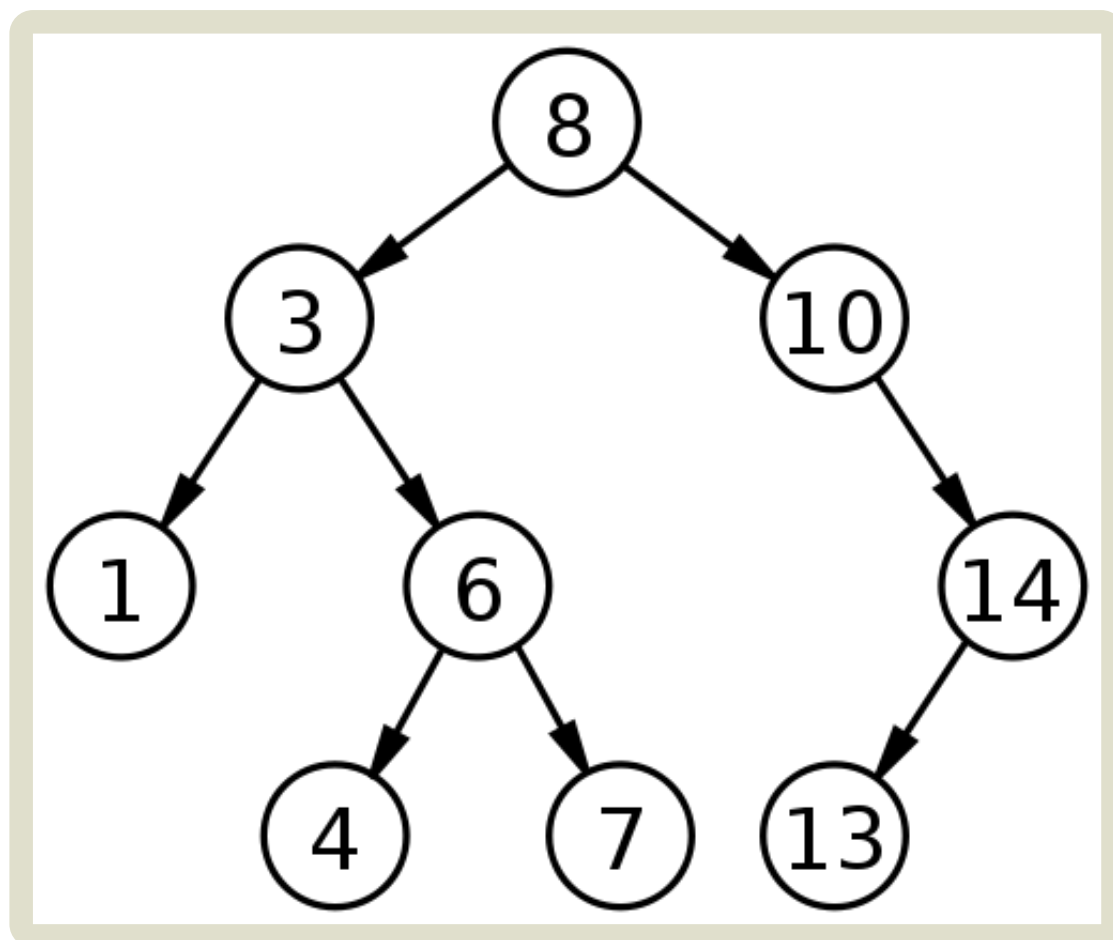
第一步，就是将所要保存的数据，写入文本文件。这个文本文件就是你的数据库。

为了方便读取，数据必须分成记录，每一条记录的长度规定为等长。比如，假定每条记录的长度是800字节，那么第5条记录的开始位置就在3200字节。

大多数时候，我们不知道某一条记录在第几个位置，只知道主键（primary key）的值。这时为了读取数据，可以一条条比对记录。但是这样做效率太低，实际应用中，数据库往往采用B树（B-tree）格式储存数据。

二、什么是B树？

要理解B树，必须从二叉查找树（Binary search tree）讲起。



二叉查找树是一种查找效率非常高的数据结构，它有三个特点。

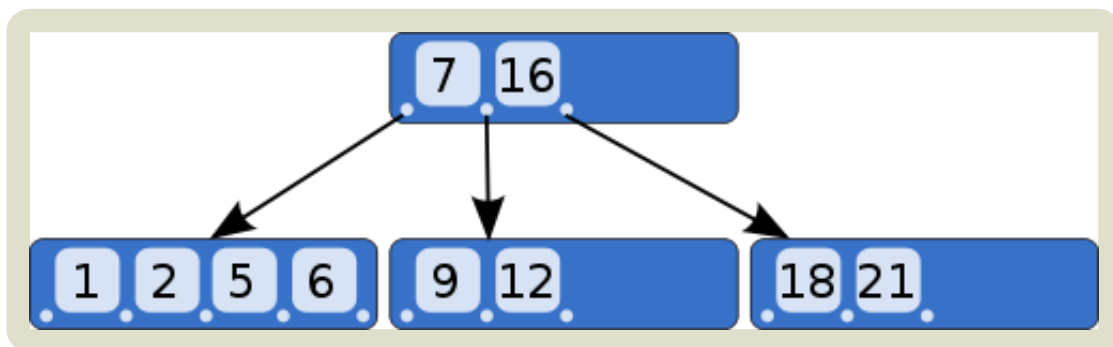
（1）每个节点最多只有两个子树。

(2) 左子树都为小于父节点的值，右子树都为大于父节点的值。

(3) 在 n 个节点中找到目标值，一般只需要 $\log(n)$ 次比较。

二叉查找树的结构不适合数据库，因为它的查找效率与层数相关。越处在下层的数据，就需要越多次比较。极端情况下， n 个数据需要 n 次比较才能找到目标值。对于数据库来说，每进入一层，就要从硬盘读取一次数据，这非常致命，因为硬盘的读取时间远远大于数据处理时间，数据库读取硬盘的次数越少越好。

B树是对二叉查找树的改进。它的设计思想是，将相关数据尽量集中在一起，以便一次读取多个数据，减少硬盘操作次数。



B树的特点也有三个。

(1) 一个节点可以容纳多个值。比如上图中，最多的一个节点容纳了4个值。

(2) 除非数据已经填满，否则不会增加新的层。也就是说，B树追求“层”越少越好。

(3) 子节点中的值，与父节点中的值，有严格的大小对应关系。一般来说，如果父节点有 a 个值，那么就有 $a+1$ 个子节点。比如上图中，父节点有两个值（7和16），就对应三个子节点，第一个子节点都是小于7的值，最后一个子节点都是大于16的值，中间子节点就是7和16之间的值。

这种数据结构，非常有利于减少读取硬盘的次数。假定一个节点可以容纳100个值，那么3层的B树可以容纳100万个数据，如果换成二叉查找树，则需要20层！假定操作系统

一次读取一个节点，并且根节点保留在内存中，那么B树在100万个数据中查找目标值，只需要读取两次硬盘。

三、索引

数据库以B树格式储存，只解决了按照"主键"查找数据的问题。如果想查找其他字段，就需要建立索引（index）。

所谓索引，就是以某个字段为关键字的B树文件。假定有一张"雇员表"，包含了员工号（主键）和姓名两个字段。可以对姓名建立索引文件，该文件以B树格式对姓名进行储存，每个姓名后面是其在数据库中的位置（即第几条记录）。查找姓名的时候，先从索引中找到对应第几条记录，然后再从表格中读取。

这种索引查找方法，叫做"索引顺序存取方法"（Indexed Sequential Access Method），缩写为ISAM。它已经有多种实现（比如C-ISAM库和D-ISAM库），只要使用这些代码库，就能自己写一个最简单的数据库。

四、高级功能

部署了最基本的数据存取（包括索引）以后，还可以实现一些高级功能。

（1）**SQL语言**是数据库通用操作语言，所以需要有一个SQL解析器，将SQL命令解析为对应的ISAM操作。







（2）**数据库连接（join）**是指数据库的两张表通过"外键"，建立连接关系。你需要对这种操作进行优化。

（3）**数据库事务（transaction）**是指批量进行一系列数据库操作，只要有一步不成功，整个操作都不成功。所以需要有一个"操作日志"，以便失败时对操作进行回滚。

（4）**备份机制**：保存数据库的副本。

（5）**远程操作**：使得用户可以在不同的机器上，通过TCP/IP协议操作数据库。

（完）

- 版权声明：自由转载-非商用-非衍生-保持署名（创意共享3.0许可证）
- 发表日期：2014年7月 4日
- 更多内容： 档案 »  理解计算机
- 付费支持： 购买文集
- 社交媒体： twitter,  weibo
- Feed订阅：

相关文章

- 2013.11.29: [Stack的三种含义](#)

学习编程的时候，经常会看到stack这个词，它的中文名字叫做"栈"。

- 2013.10.14: [为什么寄存器比内存快？](#)

计算机的存储层次（memory hierarchy）之中，寄存器（register）最快，内存其次，最慢的是硬盘。

- 2013.08.17: [Linux 的启动流程](#)

半年前，我写了《计算机是如何启动的？》，探讨BIOS和主引导记录的作用。

- 2013.07.04: [RSA算法原理（二）](#)

上一次，我介绍了一些数论知识。

广告（购买广告位）



留言（23条）

Nuk 说：

没有搞清楚索引的意思。

索引的意思是说，先按照姓名的值在B树中进行查找，找到姓名的索引号。再按照姓名的索引号在雇员表中找到雇员的所有信息吗？这和把姓名作为主键进行查找有何区别？

[2014年7月 4日 15:19](#) | [档案](#) | [引用](#)

阮一峰 说：

@Nuk：

改写了几句话，现在应该好懂一些了吧。

[2014年7月 4日 15:37](#) | [档案](#) | [引用](#)

碧浪飞虹 说：

笔误：

一般来说，如果父节点有a个值，那么就有n+1个子节点。

应为：

一般来说，如果父节点有a个值，那么就有a+1个子节点。

或者：

一般来说，如果父节点有n个值，那么就有n+1个子节点。

[2014年7月 4日 15:44](#) | [档案](#) | [引用](#)

John 说:

transaction的翻译应该是事务不是交易...

[2014年7月 4日 16:16](#) | [档案](#) | [引用](#)

土木坛子 说:

在RSS里看到有几个错字，本想来报告的，没想到已经更正了。

[2014年7月 4日 16:32](#) | [档案](#) | [引用](#)

simplejoy 说:

这样实现后，就是一个sqlite了

[2014年7月 4日 16:46](#) | [档案](#) | [引用](#)

tzp 1991 说:

好厉害的样子。连着看了几篇文章写的都很好。

[2014年7月 4日 16:56](#) | [档案](#) | [引用](#)

nickey 说:

这是关系型数据库的，非关系型可以更简单

[2014年7月 4日 17:18](#) | [档案](#) | [引用](#)

阮一峰 说:

@碧浪飞虹, @John:

谢谢指出，已经改过来了。

[2014年7月 4日 18:16](#) | [档案](#) | [引用](#)

afu1982 说:

>二叉查找树是一种查找效率非常高的数据结构，它有两个特点。

应该是三个特点吧？

[2014年7月 4日 21:05](#) | [档案](#) | [引用](#)

阮一峰 说：

引用afu1982的发言：

应该是三个特点吧？

谢谢指出，已经更正了。

[2014年7月 4日 21:48](#) | [档案](#) | [引用](#)

某人 说：

确实挺有意思，不过自己写个数据库有必要吗，或者在什么情况下，我们应该自己写个数据库，我是一个ERP实施顾问，工作中主要用到的就是数据库。都是商业级的。oracle要比sqlserver好很多。母鸡为什么

[2014年7月 4日 23:34](#) | [档案](#) | [引用](#)

何朝城 说：

引用土木坛子的发言：

在RSS里看到有几个错字，本来来报告的，没想到已经更正了。

刚过来就看到更正了(我这是不是在找别人的茬?而忘了自己)

[2014年7月 5日 00:24](#) | [档案](#) | [引用](#)

hzyy 说：

引用nickey的发言：

这是关系型数据库的，非关系型可以更简单

同意，非关系型的可能更简单，例如Google的leveldb，是一个较为简单的键值对型数据库，写得非常好

[2014年7月 5日 11:26](#) | [档案](#) | [引用](#)

xsc 说：

真是好文章，最近正好要自己实现一个微型数据库，无从下手，搜索引擎搜索出来的好多资料都很扯，每次读您的文章都大有收获，感觉您比我的大多数专业计算机老师都强的多，

[2014年7月 6日 04:00](#) | [档案](#) | [引用](#)

SY 说：

好文章！问个问题

"每个姓名后面是其在数据库中的位置（即第几条记录）" 这里的第几条记录是索引吗？

[2014年7月 7日 09:36](#) | [档案](#) | [引用](#)

xjhns 说：

"在n个节点中找到目标值，一般只需要 $\log(n)$ 次比较。"

该怎么理解，是最少需要 $\log(n)$ 次比较吗？

[2014年7月 7日 10:43](#) | [档案](#) | [引用](#)

breezefeng 说：

要不要实现一个呢？给个具体例子呗

[2014年7月 8日 13:08](#) | [档案](#) | [引用](#)

hello 说：

这篇写得有点外行了。

首先，数据库 有很多种，这里其实想说的是 mysql oracle这种关系数据库的实现。

但是B+树其实只是一种查找数据的数据结构，和任何一种数据库没有必然的关系，

使用这种数据结构的不一定就是关系数据库。

我觉得要说实现了一个关系数据库的底线是描述这种实现，对于从第一范式到BCNF的符合度。

[2014年7月 8日 13:40](#) | [档案](#) | [引用](#)

查sir 说：

引用Nuk的发言：

没有搞清楚索引的意思。索引的意思是说，先按照姓名的值在B树中进行查找，找到姓名的索引号。再按照姓名的索引号在雇员表中找到雇员的所有信息吗？这和把姓名作为主键进行查找有何区别？

关键是姓名不是主键啊

[2014年7月 9日 11:10](#) | [档案](#) | [引用](#)

LYJ 说：

引用hello的发言：

这篇写得有点外行了。

首先，数据库 有很多种，这里其实想说的是 mysql oracle这种关系数据库的实现。

但是B+树其实只是一种查找数据的数据结构，和任何一种数据库没有必然的关系，使用这种数据结构的不一定就是关系数据库。

我觉得要说实现了一个关系数据库的底线是描述这种实现，对于从第一范式到BCNF的符合度。

我倒是觉得范式也未必体现关系数据库本质吧，更加本质的应该是基于数学的关系

演算，是集合论的一种体现。关系数据库因为用严格的数学推导证明，所有更合适存储关键的，重要的数据，这应该的和文档数据库等其他数据库不同的地方。

[2014年7月12日 18:26](#) | [档案](#) | [引用](#)

introoom 说：

引用Nuk的发言：

没有搞清楚索引的意思。

索引的意思是说，先按照姓名的值在B树中进行查找，找到姓名的索引号。

再按照姓名的索引号在雇员表中找到雇员的所有信息吗？这 and 把姓名作为主键进行查找有何区别？

Hi, Nuk. 正好看到， お久しぶり。IIUC，姓名不可以做主键，数据是按照姓名排序，所以index姓名，得到第一个entry地址，然后开始sequential读取。

[2014年7月21日 16:07](#) | [档案](#) | [引用](#)

zhengjunwei 说：

不错啊，mysql就有一个MYISAM(Indexed Sequential Access Method)，这么说用的是B树哈

[2014年7月22日 01:02](#) | [档案](#) | [引用](#)

我要发表看法

您的留言（HTML标签部分可用）

您的大名：

«-必填

电子邮件：

«-必填，不公开

个人网址：

«-我信任你，不会填写广告链接

记住个人信息？ ☐

发表

«- 点击按钮