空间数据库（5）

陈斌
gischen@pku.edu.cn
2015.11.05

空间存储和索引

- 目标和基本思想
- 物理存储介质
- 缓冲区管理
- 存储组织
- 存取路径：索引结构

目标和基本思想

- 概念和逻辑层次
  - 空间概念模型
  - 空间查询语言
- 物理存储与数据结构
  - 物理存储器分级管理
  - 存储组织
  - 聚类和空间聚类
  - 索引和空间索引

目标和基本思想

- 存储器分级管理
  - 解决速度快、容量小、易失和速度慢、容量大、持久之间的矛盾
  - 考虑I/O代价，采用缓存机制
- 存储组织
  - 适合外存操作的数据结构
  - 记录和记录的多种组织形式

目标和基本思想

- 聚类和空间聚类
  - 降低大查询的磁盘寻道时间和等待时间
  - 空间上相邻的对象在外存中也相邻
- 索引和空间索引
  - 提供多种存取路径，提高查询效率
  - 高效的排序树结构及其空间扩展

数据库存储管理

- 如何存储和管理海量的数据？
  - 目标数据、元数据、索引、日志等
- 采用何种表示方式和数据结构对数据处理提供最佳、最有效的支持？

物理存储介质：基本存储

- 寄存器（register）
  - CPU的一部分，用于暂存运算中间结果
  - 与运算部件直接连接，速度最快，极少（几十个）
- 高速缓冲存储器（cache memory）
  - CPU的一部分，用于缓存主存储器
  - 在CPU中，速度极快，容量小（几十K~几M）
  - 操作系统底层管理

物理存储介质：基本存储

- 主存储器（main memory）
  - 通过总线与CPU相连，存储运算所需的数据和指令
  - 速度很快（纳秒级），一般容量在几百M～几G
  - 随机访问：访问任何存储单元时间相同
  - 易失性：断电丢失
  - 操作系统提供机制，应用程序管理

物理存储介质：在线存储

- 快闪存储器（flash memory）
  - 通过外设接口与总线相连，存储永久保留的数据
  - 速度受到存储介质和接口限制
  - 随机访问，非易失性，断电不丢失
  - 文件系统管理，可以通过操作系统在线访问
- 磁盘存储器（disk memory）
  - 同上，但是机械装置，速度更慢

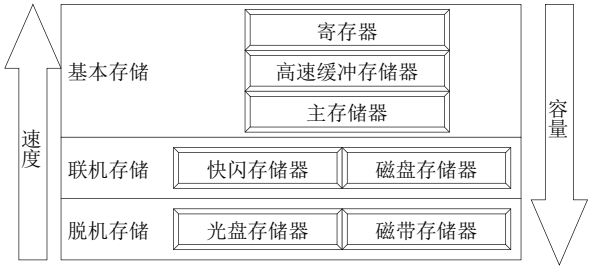
物理存储介质：脱机存储

- 光盘存储器（CDROM/CDR/CDRW/DVD）
  - 脱机存储，保存备份或者历史档案
  - 分为光物理盘，光化学盘和光磁盘
  - 只读，可写一次，可重复读写
  - 机械装置，随机访问，速度更低
  - 有标准数据记录格式，操作系统提供文件系统接口访问

物理存储介质：脱机存储

- 磁带存储器（tape）
  - 脱机存储，保存备份或者历史档案
  - 电磁记录原理
  - 机械装置，顺序访问
  - 速度最低，容量价格比最高（至几百G）

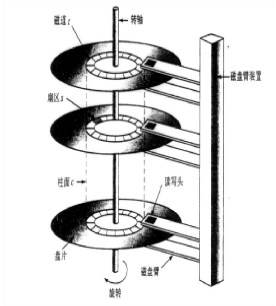
物理存储介质层次



磁盘存储器

- 磁盘物理特性
- 磁盘性能度量
- 磁盘块存取的优化
- 磁盘阵列技术RAID

磁盘存储物理特性



- 盘片
  - 磁道track
  - 扇区sector
  - 柱面cylinder
- 磁头
- 驱动臂

磁盘存储器性能度量

- 容量：磁盘所能容纳的字节数
- 存取时间：从发出读写请求到数据开始传输之间的时间
  - 寻道时间：移动磁盘臂，定位到正确磁道所需的时间。（平均4 - 10毫秒）
  - 旋转等待时间：等待被存取的扇区出现在读写头下所需的时间。（平均2 - 5毫秒）
- 数据传输率：从磁盘获得数据或者向磁盘存储数据的速率
- 平均故障时间（可靠性）：预期系统无故障连续运行的平均时间MTBF

磁盘块存取的优化

- 在主存储器中对块进行缓冲以减少块的读写次数
- 按柱面组织数据
- 磁盘臂调度 -- 电梯算法
- 利用非易失性RAM作为写缓冲
- 预读和双缓冲

磁盘阵列技术RAID

- RAID: Redundant Array of Inexpensive(Independent) Disks
  - 通过冗余提高可靠性
  - 通过并发访问提高性能
- 冗余方案
  - 镜像：复制写到多个磁盘
  - 校验位：写入数据及其校验和
  - 损失有效容量

磁盘阵列技术RAID

- 并发方案
  - 位级拆分：将每个字节按位分开，存储到多个磁盘上
  - 块级拆分：对块进行逻辑编号，对于n个磁盘的阵列，将逻辑上的第i块存储到第（i mod n）+1个磁盘上

磁盘阵列技术RAID

- 标准RAID级别
  - RAID 0级：块级拆分，无冗余
  - RAID 1级：带块级拆分的磁盘镜像
  - RAID 2级：内存风格的纠错码组织结构
  - RAID 3级：位交叉的奇偶校验组织结构
  - RAID 4级：块交叉的奇偶校验组织结构
  - RAID 5级：块交叉的分布奇偶校验位的组织结构
  - RAID 6级：P+Q冗余方案

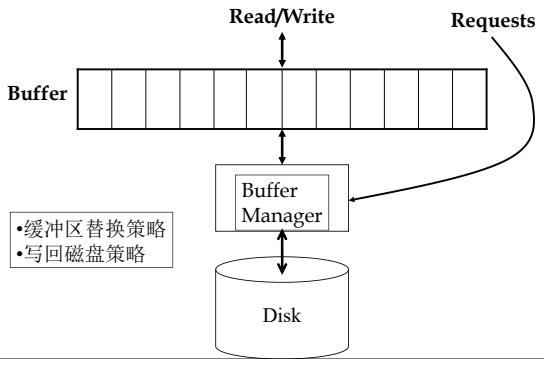
磁盘阵列技术RAID

- RAID级别的选择
  - 所需的额外磁盘存储带来的开销
  - 在I/O操作数量方面的性能需求
  - 磁盘故障时的性能
  - 数据重建过程中的性能

缓冲区管理

- 缓冲区buffer
  - 主存储器中用于存储磁盘块的拷贝的部分，由固定数目的缓冲块构成
- 缓冲区管理器
  - 负责缓冲区空间分配调度的子系统
- 缓冲区管理的目的
  - 减少磁盘和主存储器之间传输的块的数目

缓冲区管理基本流程



缓冲区替换策略

- 最近最少使用（LRU）
  - 替换出最长时间没有读或写过的块
- 先进先出（FIFO）
  - 替换出被同一个块占用时间最长的缓冲块
- “时钟”算法
  - LRU的一个常见的、有效的近似
- 系统控制
  - 查询优化器或者其它的DBMS部件可以给缓冲区管理器提供建议来避免象LRU，FIFO，或者时钟这样的严格的策略可能引起的问题

最近最少使用（LRU）

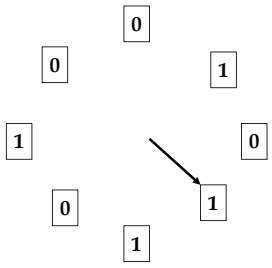
- 基本方法
  - 缓冲区管理器保持一个表，登记每个缓冲块被访问的最后一次时间
  - 每个数据库访问在这个表中生成一个表项。
- LRU是一个有效的策略
  - 直觉上，长时间没有使用的缓冲区比那些最近访问过的缓冲区有更小的最近访问的可能性
  - 但也有例外

先进先出（FIFO）

- 基本方法
  - 缓冲区管理器保持一个表，登记当前占用缓冲区的各个块装入缓冲区的时间
  - 当块从磁盘读入内存的时候，生成一个表项
  - 当块被访问时，不需要修改这个表
- 与LRU相比，FIFO需要较少的维护
  - 但它存在更多的问题
  - 例如被重复使用的，B-树索引的根块将最终变成一个缓冲区中最旧的块它将被写回到磁盘上，很快又被重新读入另一个缓冲区。

“时钟”算法

- 基本方法
  - 将缓冲区看作一个环。一个指针指向一个缓冲块
  - 每一个缓冲块有一个“标志”，0或1
  - 带有0标志的是很可能被替换出去的缓冲块
  - 刚读入和访问过的块标记为1
  - 指针旋转过程中不停将1变为0



缓冲区管理系统控制

- 查询优化器或者其它的DBMS部件可以给缓冲区管理器提供建议
  - 例如，将某些块定义为“固定的”来强迫它们保持在内存中，如B-树的根、数据字典中的块。
  - 又如，对于象一遍散列连接那样的算法，查询处理器可以“固定”较小的关系的块，使得确保在全部时间内它都将留在内存中。

块写回控制策略

- 被钉住的块
  - 为了使数据库系统能够从崩溃中恢复，必须限制一个块写回磁盘的时间。不允许写回磁盘的块称为被钉住的块
- 块的强制写出
  - 某些情况下，尽管不需要一个块所占用的缓冲区空间，仍必须把这个块写回磁盘。这样的写操作称为块的强制写出。

数据存储组织

- 数据库中的数据存储在操作系统管理的文件中
  - 域->记录->(块)->文件
- 域根据类型占据不同大小空间
  - 定长域类型
  - 变长域类型
  - 二进制大对象类型（BLOB）
    - 常用于空间复杂对象的存储，至少可以提供存储管理和事物支持

数据存储组织

- 记录由域顺序排列组成
  - 定长记录
  - 变长记录：含有变长域的记录
- 定长记录文件
  - 文件中所有的记录都具有相同的长度，从而一个块中所有的记录都是等长的
- 变长记录文件
  - 文件中的记录可以有不同的长度，从而一个块中的各个记录可以具有不同的长度

数据存储组织

- 定长记录文件的顺序存储
  - 块的大小应为记录大小的整倍数
    - 对齐
  - 删除记录代价高
    - 删除标记，或
    - 有效指针链接

记录0	Perryridge	A-102	400
记录1	Rouond Hill	A-305	350
记录2	Mianus	A-215	700
记录3	Downtown	A-101	500
记录4	Redwood	A-222	700
记录5	Perryridge	A-201	900
记录6	Brighton	A-217	750
记录7	Downtown	A-110	600
记录8	Perryridge	A-218	700

数据存储组织

- 变长记录文件
  - 多种记录类型在一个文件中存储
  - 记录类型允许一个或多个字段是变长的
  - 记录类型允许可重复的字段
- 变长记录文件的定长表示法
  - 预留空间：使用长度为最大记录长度的定长记录。对较短记录未使用的空间用特殊的空值或记录终结符号来填充。
  - 使用指针：变长记录用一系列通过指针链接起来的定长记录来表示。

变长记录文件：预留空间法

记录0	Perryridge	A-102	400	A-201	900	A-218	700
记录1	Round Hill	A-305	350	⊥	⊥	⊥	⊥
记录2	Mianus	A-215	700	⊥	⊥	⊥	⊥
记录3	Downtown	A-101	500	A-110	600	⊥	⊥
记录4	Redwood	A-222	700	⊥	⊥	⊥	⊥
记录5	Brighton	A-217	750	⊥	⊥	⊥	⊥

变长记录文件：使用指针

记录0	Perryridge	A-102	400	
记录1	Rouond Hill	A-305	350	
记录2	Mianus	A-215	700	
记录3	Downtown	A-101	500	
记录4	Redwood	A-222	700	
记录5		A-201	900	
记录6	Brighton	A-217	750	
记录7		A-110	600	
记录8		A-218	700	

变长记录文件：锚块和溢出块

锚块

Perryridge	A-102	400	
Round Hill	A-305	350	
Mianus	A-215	700	
Downtown	A-101	500	
Redwood	A-222	700	
Brighton	A-217	750	

溢出块

A-210	900	
A-218	700	
A-110	600	

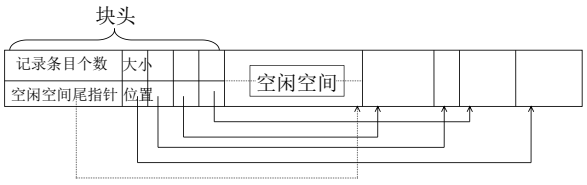
变长记录文件：字节流表示法

- 字节流表示：把每个记录作为一个连续的字节流存储，每个记录的末尾附加一个特殊的记录终止符号

0	Perryridge	A-102	400	A-201	900	A-218	700	⊥
1	Round Hill	A-305	350	⊥				
2	Mianus	A-215	700	⊥				
3	Downtown	A-101	500	A-110	600	⊥		
4	Redwood	A-222	700	⊥				
5	Brighton	A-217	750	⊥				

变长记录文件：分槽的页结构

- 分槽的页结构：每个块的开始处有一个块头



数据存储组织

- 文件中记录的组织
  - 关系中的各个记录存放在文件中的什么位置
  - 堆文件组织：记录没有顺序,一条记录可以放在文件中的任何地方。
  - 散列文件组织：散列函数的计算结果确定记录应存储到文件的哪个块中。
  - 顺序文件组织：记录根据搜索码的值顺序存储

数据字典的存储

- 数据字典：数据库的描述信息
  - 关系模式信息：逻辑结构
  - 关系存储信息：物理结构
  - 用户信息：安全控制
  - 统计信息：数量/容量统计
  - 索引信息……
- RDBMS中，数据字典和普通关系同样存储

参考文献

- [TP311.13/261]空间数据库 = Spatial databases a tour (美) Shashi Shekhar, Sanjay Chawla著 谢昆青 ... 等译 北京 机械工业出版社 2004
- 北京大学计算机系数据库教研室
  - 数据库原理与技术讲义（杨冬青）
- 数据库系统概论
  - 岳丽华，丁卫群 编著
  - 科学出版社，2000年