第17讲 数据库物理存储

本讲学习什么?

基本内容

1. 基础回顾-计算机系统的存储体系
2. 磁盘的结构与特性
3. DBMS数据存储与查询实现的基本思想
4. 数据库之表和记录与磁盘块的映射
5. 数据库之文件组织方法?

重点与难点

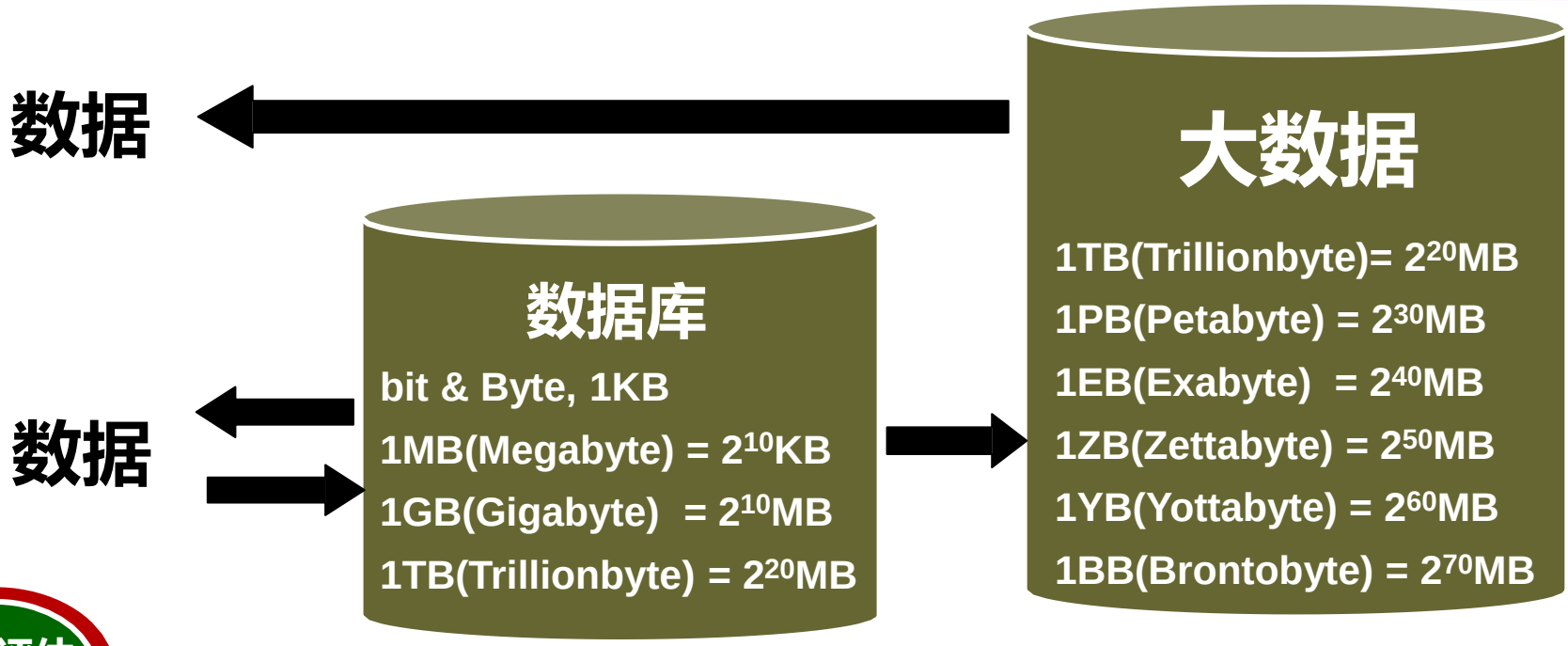
- 理解利用磁盘组织大规模数据的基本思维
- 初步了解数据存储与查询实现的基本思想
- 理解三种文件组织方法及其特性：堆文件、顺序文件和散列文件
- 理解数据库重组的概念和作用

数据库物理存储

1. 基础回顾-计算机系统的存储体系
2. 磁盘的结构与特性
3. DBMS数据存储与查询实现的基本思想
4. 数据库之表和记录与磁盘块的映射
5. 数据库之文件组织方法?

1. 基础回顾-计算机系统的存储体系

(1)数据库的存储与检索问题



性能评估与改进

既是DBA的职责—借助于软件管理与维护，又是计算机学者需要研究和解决的问题

算法设计与实现

两个基本问题如何解决?

- **如何高效率的存储? --数据组织与索引**
- **如何快速的检索? --查询实现与查询优化**

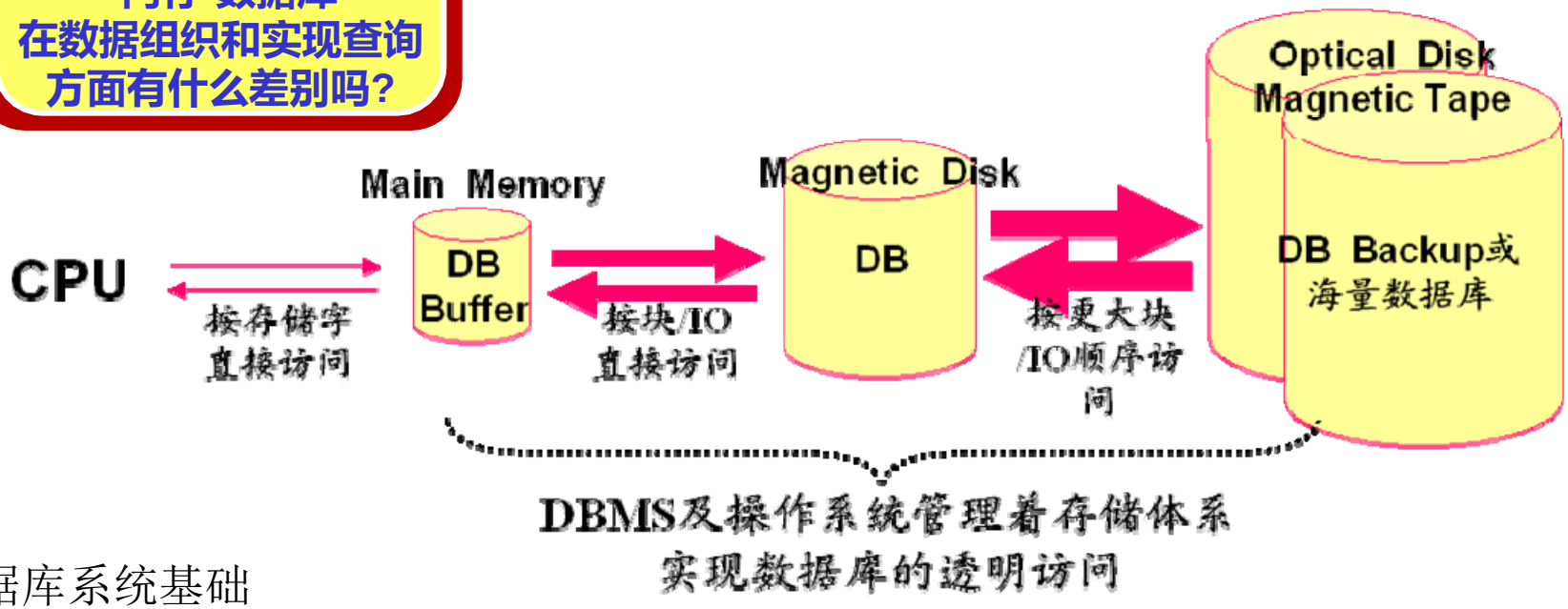
面向大规模用户, 又如何解决?

1. 基础回顾-计算机系统的存储体系

(2)什么是存储体系

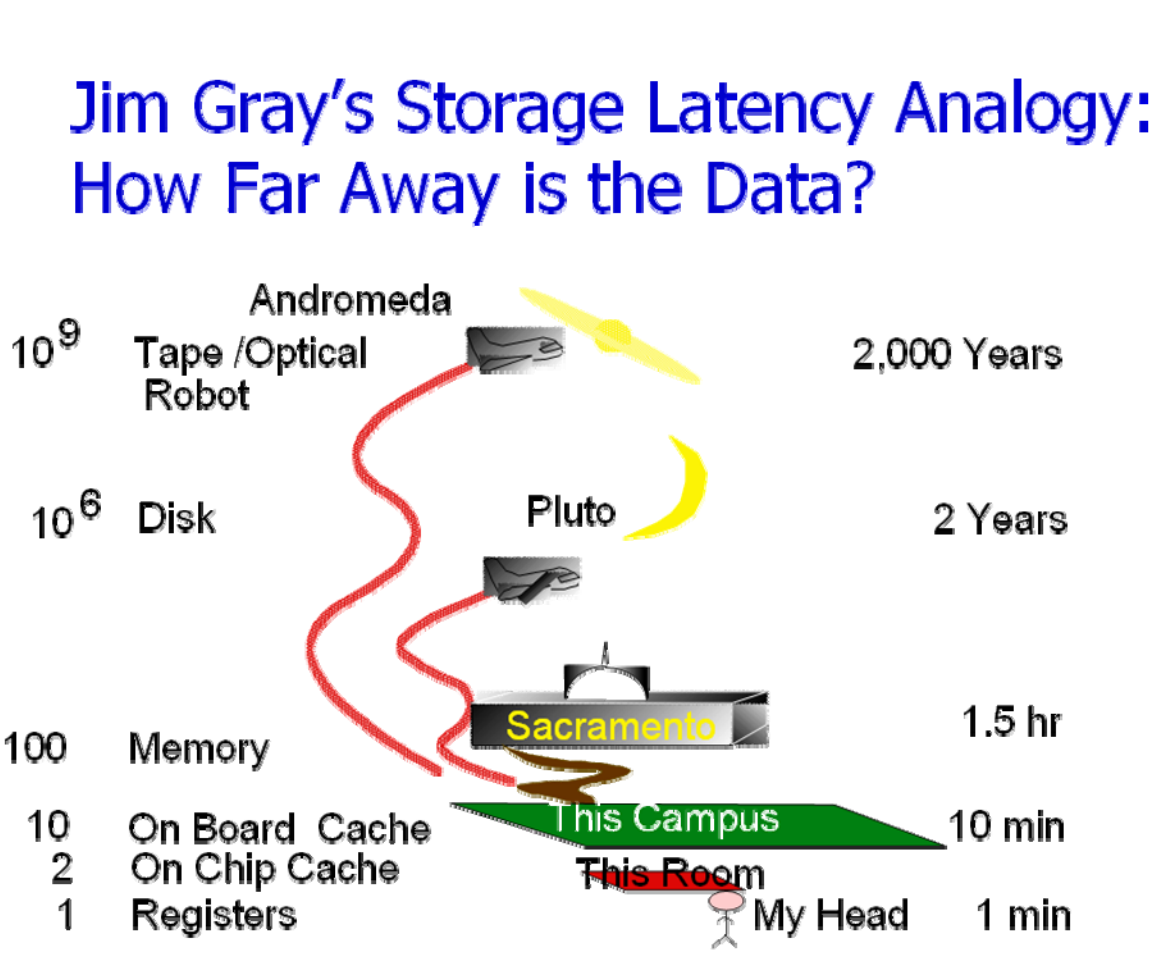
- **数据组织的基础--存储体系**
  - **将不同性价比的存储器组织在一起，满足高速度、大容量、低价格 需求**
  - **CPU与内存直接交换信息，按存储单元(存储字)进行访问**
  - **外存按存储块进行访问，其信息需先装入内存，才能被CPU处理**

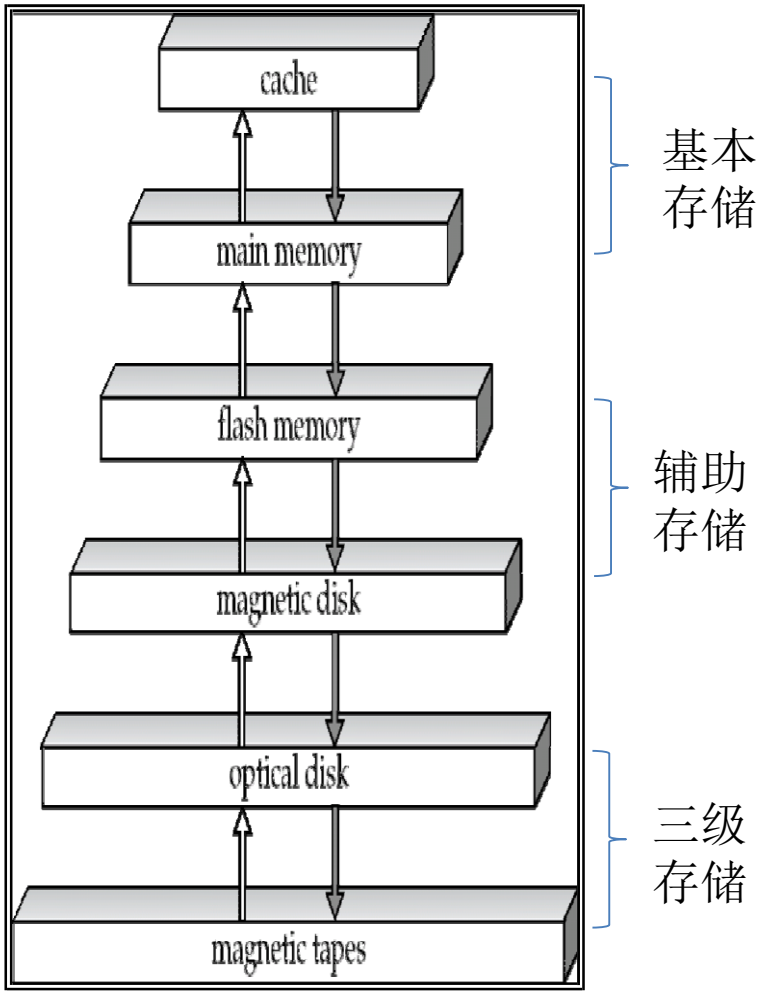
•磁带-数据库
•磁盘-数据库
•内存-数据库
在数据组织和实现查询方面有什么差别吗?



1. 基础回顾-计算机系统的存储体系

(3)不同层次存储的访问时间上的差异



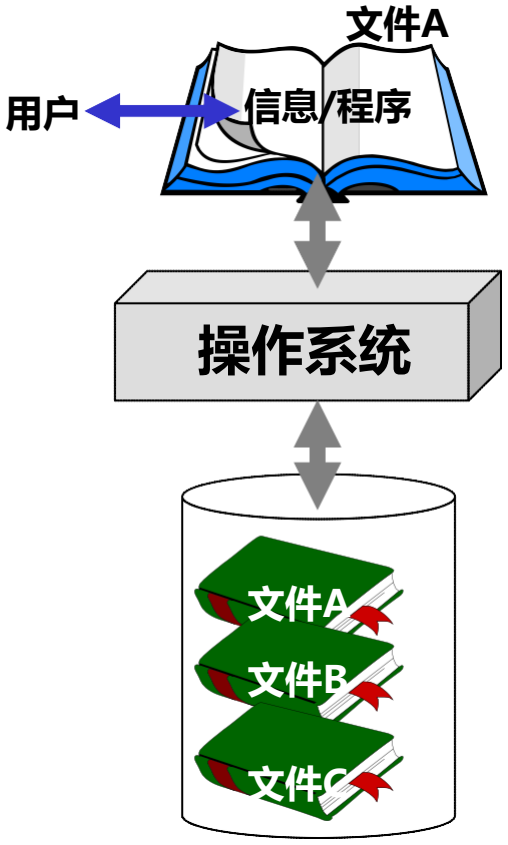


1. 基础回顾-计算机系统的存储体系

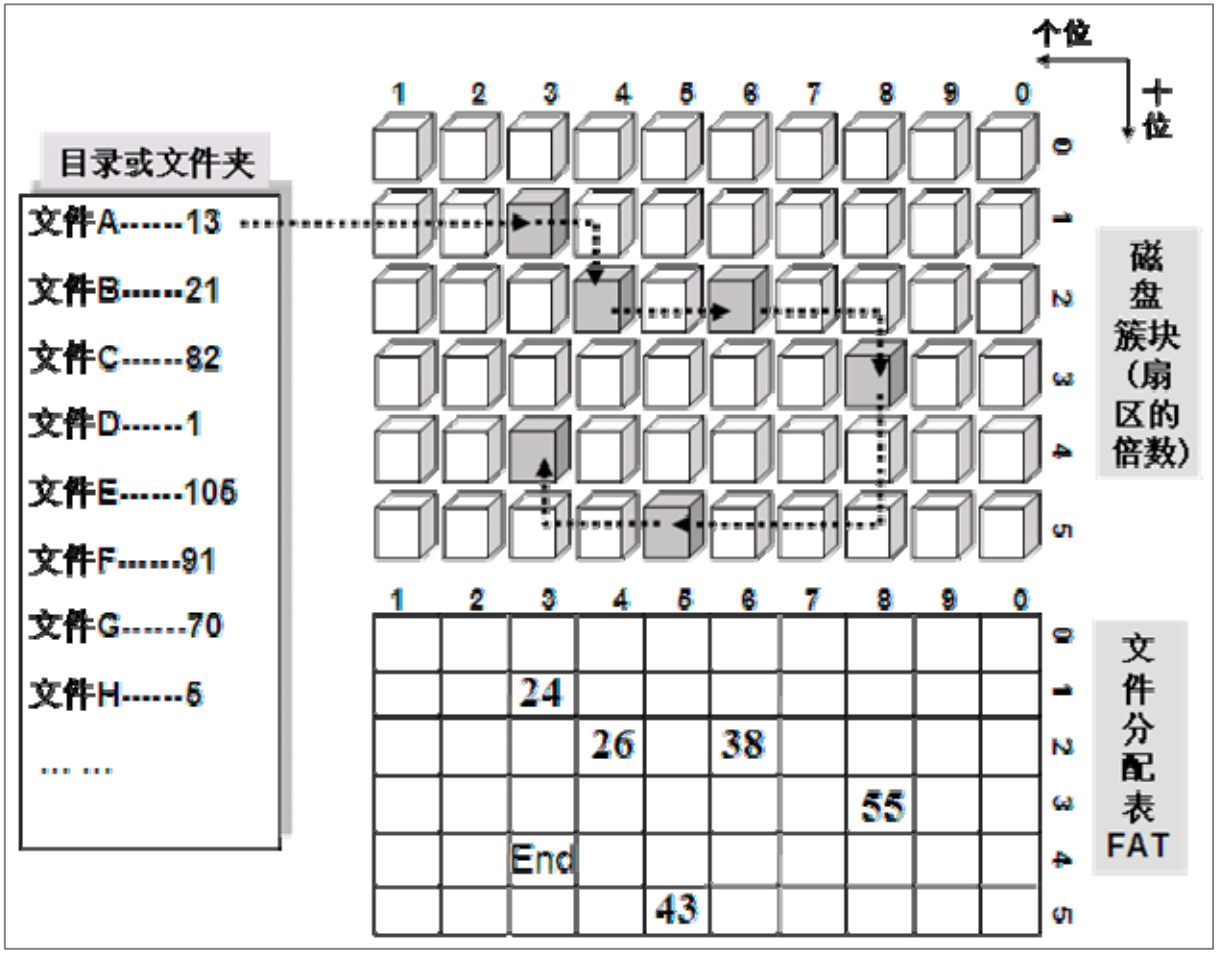
(4)操作系统如何管理磁盘和数据

操作系统对数据的组织：FAT-目录(文件夹)-磁盘块/簇

- **FAT(文件分配表-File Allocation Table)**



外存(硬盘/软盘/光盘)

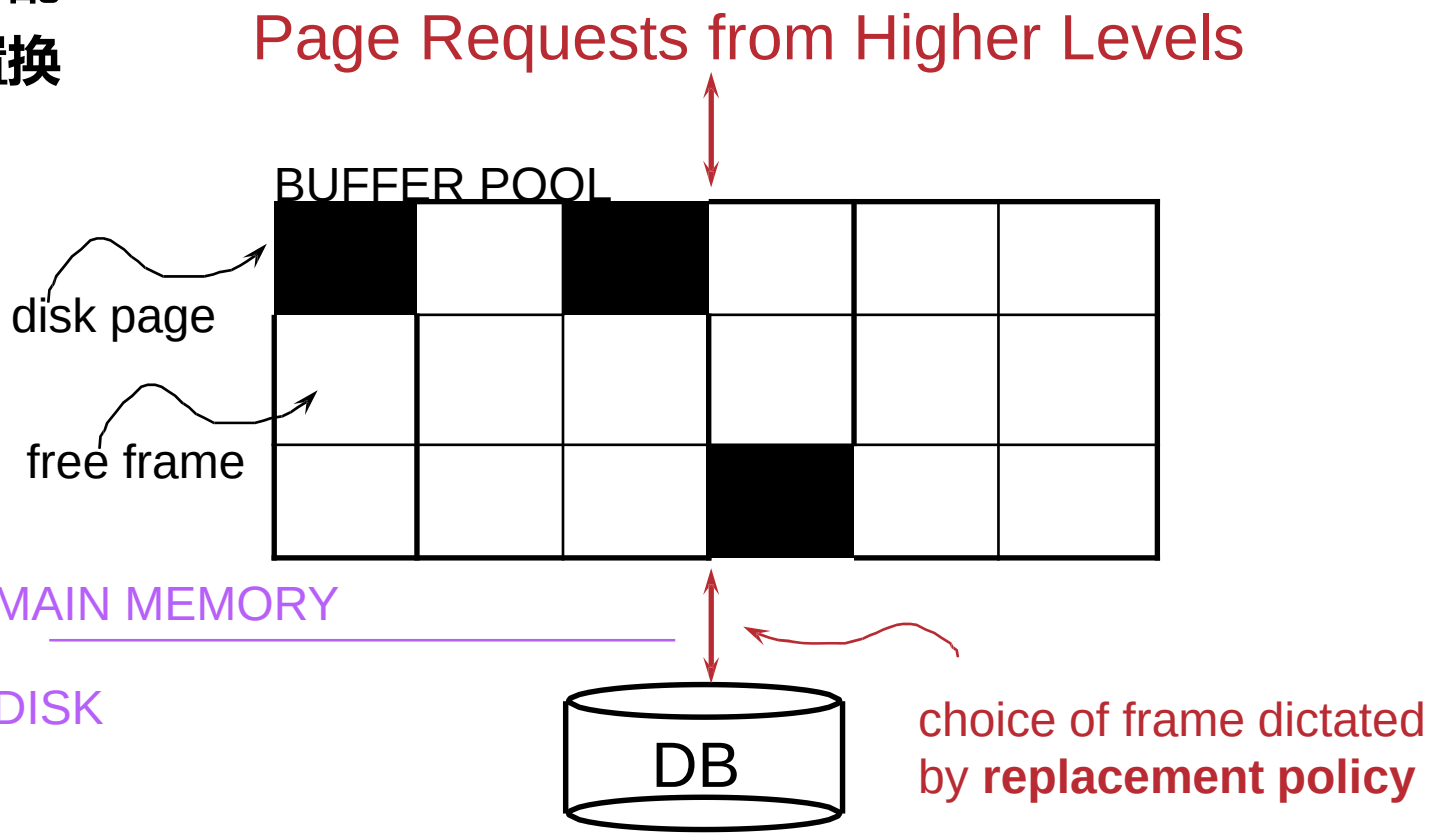


1. 基础回顾-计算机系统的存储体系

(5)操作系统对内存-缓冲区的管理

内存管理

- **一条记录的地址=存储单元的地址=内存地址=页面：页内偏移量**
- **页面(Page) = 块(Block)**
- **内存页面的分配**
- **内存页面的置换**



第17讲 数据库物理存储

1. 基础回顾-计算机系统的存储体系
2. 磁盘的结构与特性
3. DBMS数据存储与查询实现的基本思想
4. 数据库之表和记录与磁盘块的映射
5. 数据库之文件组织方法?

2. 磁盘的结构与特性

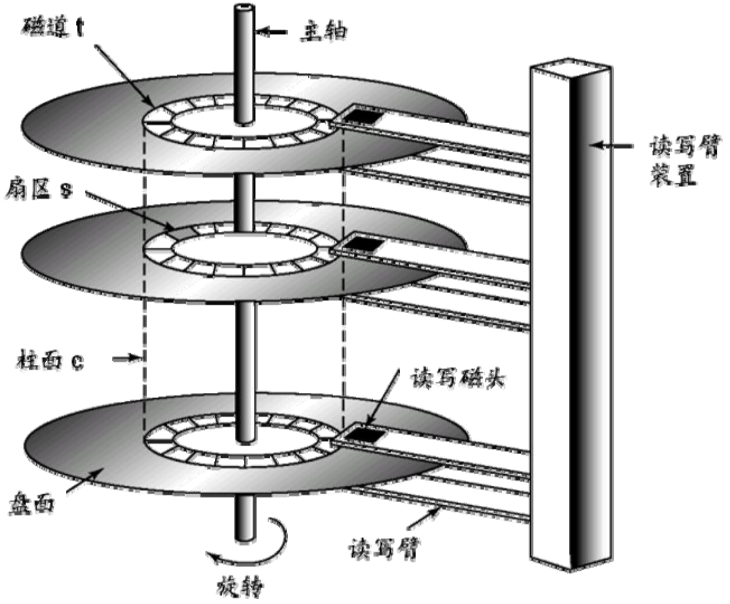
(1)磁盘及磁盘的容量

盘面: 磁道: 扇区

- **磁盘读写单位：Sector → 簇Cluster/块Block:连续的若干个扇区**

示例：一个磁盘的基本信息
- **8个圆盘，16个盘面**
- **每个盘面有2^{16}或65536个磁道**
- **每个磁道(平均)有2^8=256个扇区**
- **每个扇区有2^{12}=4096个字节**

磁盘的容量=$2^4*2^{16}*2^8*2^{12}=2^{40}$字节



2. 磁盘的结构与特性

(2)磁盘数据的访问

磁盘数据读写时间

–寻道时间(约在1-20ms)

–旋转时间(约0-10ms)

–传输时间(每4KB页<1ms)

示例：一个磁盘的基本信息

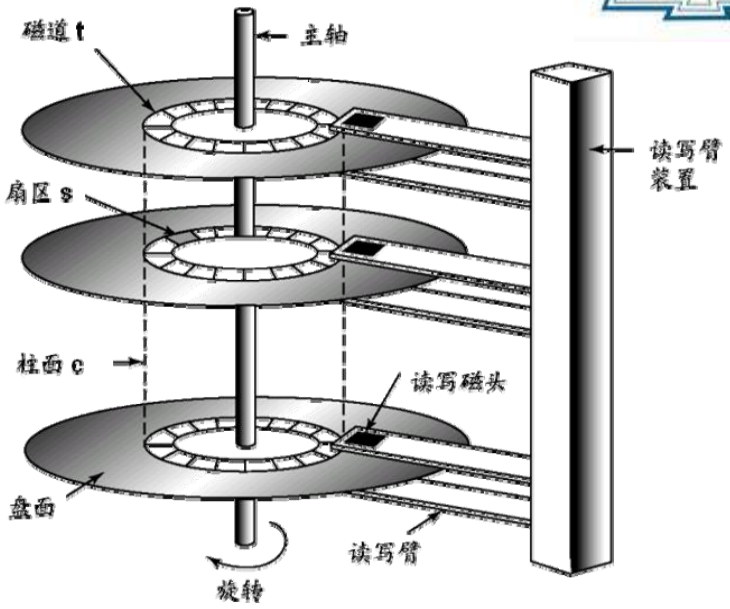
- 磁盘以7200转/min旋转，则：8.33ms内旋转一周
- 柱面之间移动磁头组合从启动到停止花费1ms,每移动4000个柱面另加1ms. 即磁头在0.00025ms内移动一个磁道，则：从最内圈移动到最外圈，移动65536个磁道大约用17.38ms.
- 一个磁道中扇区间的空隙大约占10%的空间
- 一个磁盘块 = 4个扇区 = 16384个字节

读一个磁盘块(=4个扇区16384个字节)：

最小时间=传输时间大约是0.13ms

最长时间 = 寻道时间+旋转时间+传输时间
= 17.38+8.33+0.13=25.84ms

平均时间 =6.46+4.17+0.13=10.76ms



物理存取算法考虑的关键:

–降低I/O次数

–降低排队等待时间

–降低寻道/旋转延迟时间：

- 同一磁道连续块存储；
- 同一柱面不同磁道并行块存储；
- 多个磁盘并行块存储

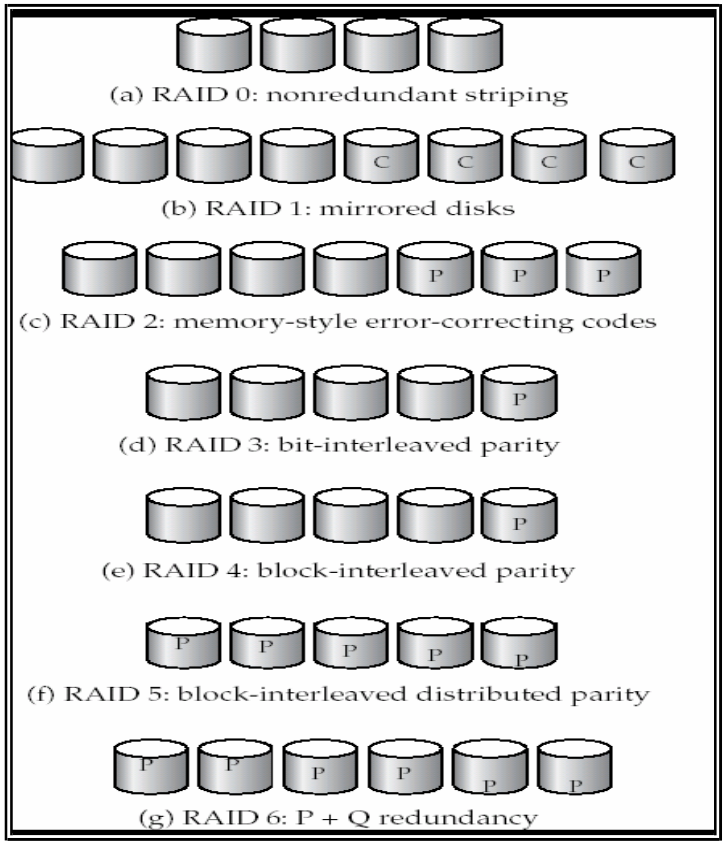
2. 磁盘的结构与特性

(3)提高磁盘数据读写时间与存储可靠性的方法

RAID技术: Redundant Array of Independent Disk

- **并行处理**: 并行读取多个磁盘
- **可靠性**: 奇偶校验与纠错

(a)**块级拆分但无冗余 raid0**
(b)**镜像处理**：每一个磁盘有一个镜像磁盘C， raid1
(c)**位交叉纠错处理**：4个磁盘存储4位+3个校验盘P存储3校验位, raid 2
(d)**位交叉校验**：4个磁盘存储4位+1个校验盘存储1校验位，位拆分存储(借助于扇区读写校验判断出错磁盘，再依据校验盘进行纠错) , raid 3
(e)**块交叉校验**：块拆分存储，其他同, raid 4
(f)**块交叉分布式校验**：块拆分存储，互为校验盘 , raid 5
(g)**更为复杂的冗余处理**(略)



- **比特级拆分**: 一个字节被拆分成8个比特位, 不同比特位存储于不同磁盘.
- **块级拆分:** 一个文件由多个块组成，不同块存储于不同磁盘.

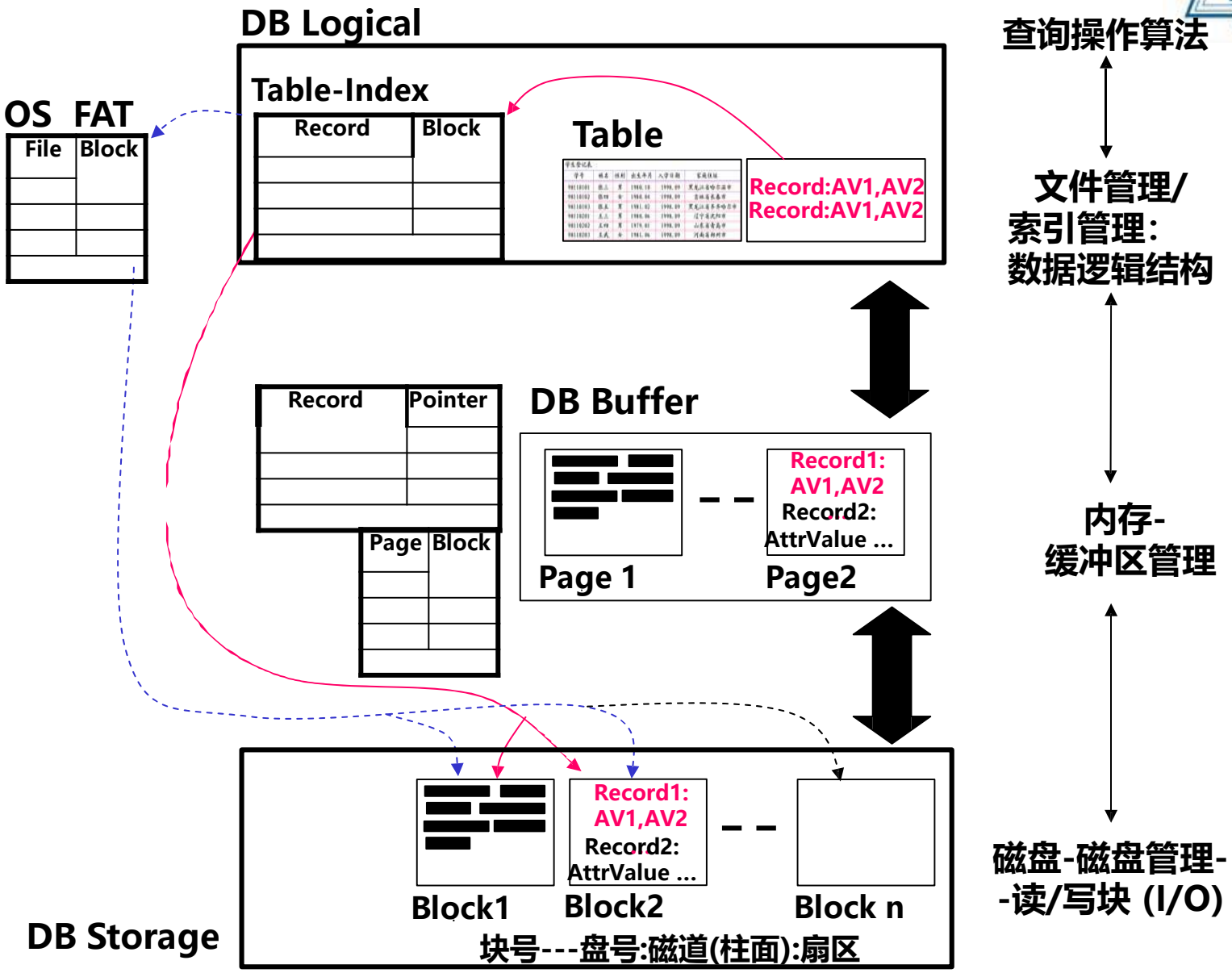
- **扇区/块读写校验**：对一个扇区/块读写做校验.
- **磁盘间读写校验:** 多个磁盘间共同构成的信息读写做校验.

第17讲 数据库物理存储

1. 基础回顾-计算机系统的存储体系
2. 磁盘的结构与特性
3. DBMS数据存储与查询实现的基本思想
4. 数据库之表和记录与磁盘块的映射
5. 数据库之文件组织方法?

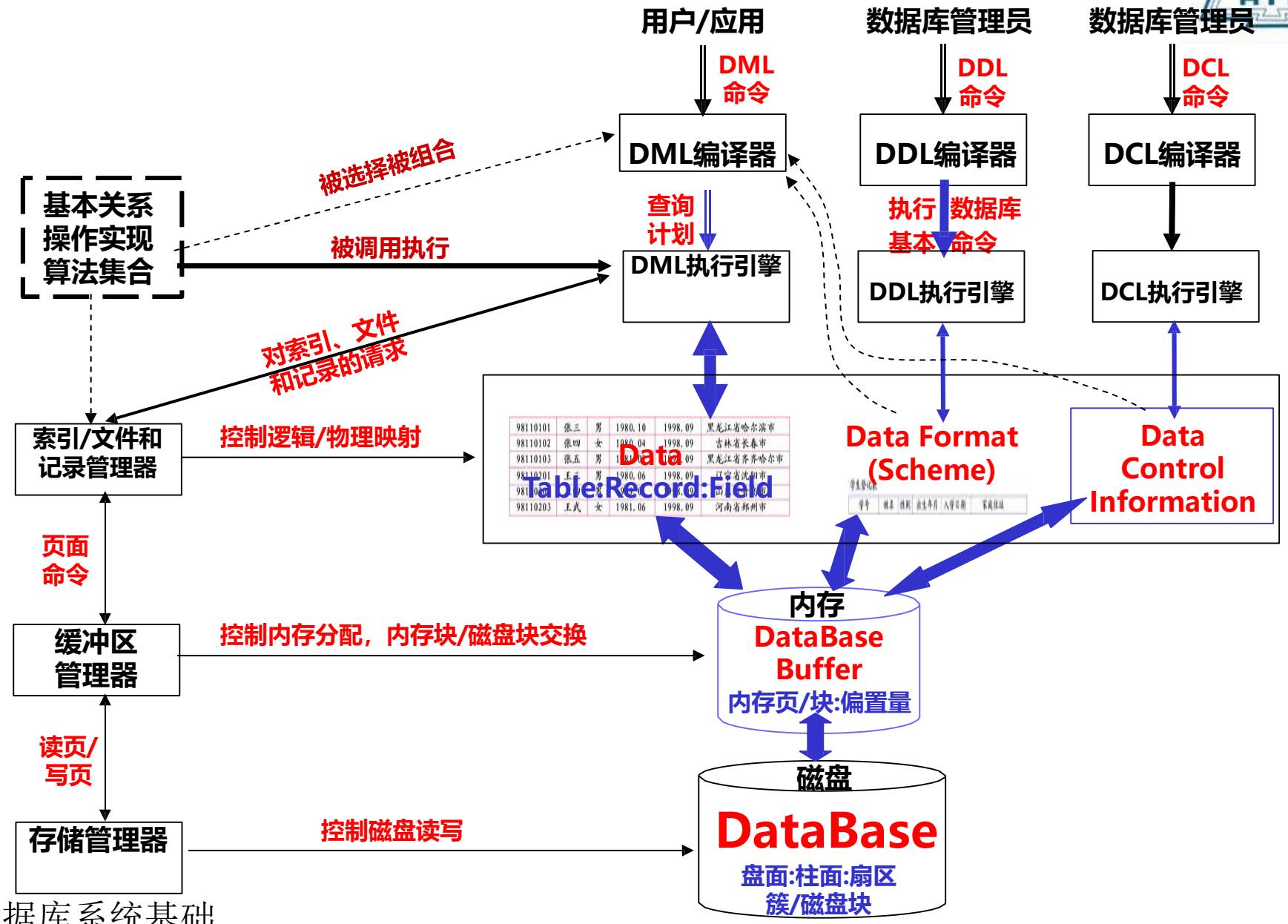
3. DBMS数据存储与查询实现的基本思想

(1)数据存储的映射关系示意



3. DBMS数据存储与查询实现的基本思想

(2)数据存储与查询实现的基本框架示意



第17讲 数据库物理存储

1. 基础回顾-计算机系统的存储体系
2. 磁盘的结构与特性
3. DBMS数据存储与查询实现的基本思想
4. 数据库之表和记录与磁盘块的映射
5. 数据库之文件组织方法?

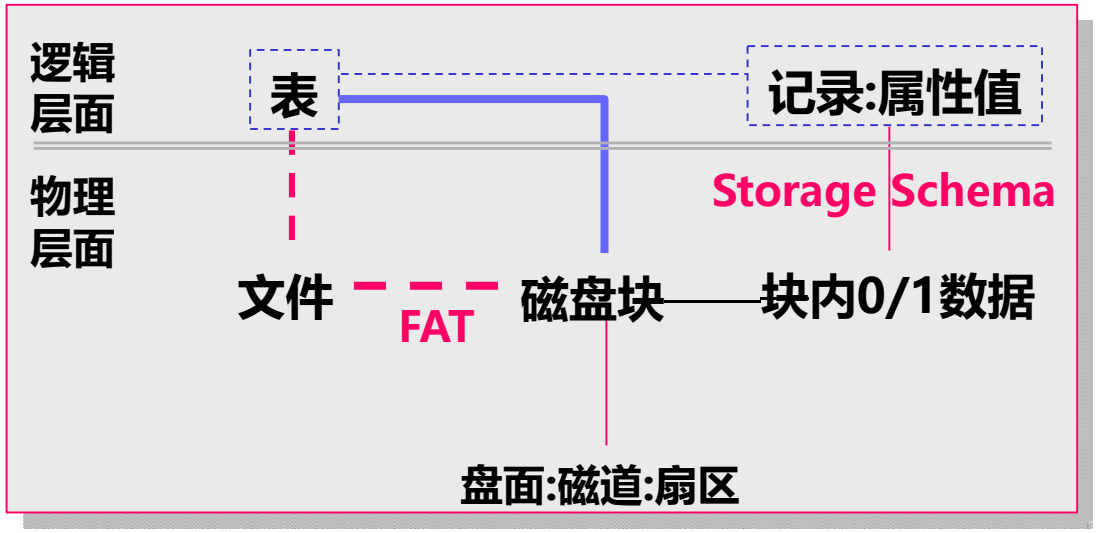
4. 数据库之表-记录与磁盘块的映射

(1)数据库概念与磁盘相关概念的映射示意

SQL—表:记录:属性值

学生登记表

学号	姓名	性别	出生年月	入学日期	家庭住址
98110101	张三	男	1980.10	1998.09	黑龙江省哈尔滨市
98110102	张四	女	1980.04	1998.09	吉林省长春市
98110103	张五	男	1981.02	1998.09	黑龙江省齐齐哈尔市
98110201	王三	男	1980.06	1998.09	辽宁省沈阳市
98110202	王四	男	1979.01	1998.09	山东省青岛市
98110203	王武	女	1981.06	1998.09	河南省郑州市



Storage—盘面:磁道:扇区

Block1

01010101010101010101001
01010010101010100101010
10011000010101 0101010

Block5

01010101010101010101001
01010010101010100101010
10011000010101 0101010

Block8

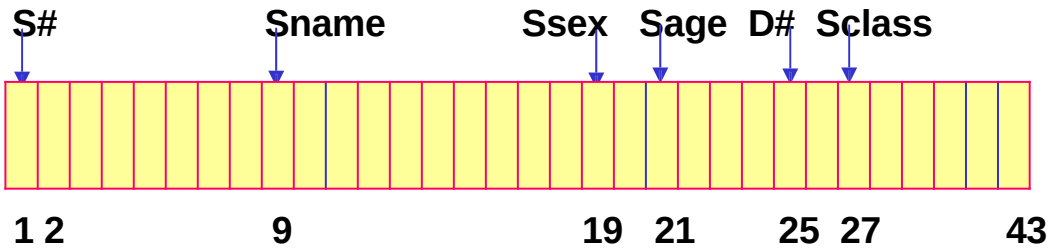
01010101010101010101001
01010010101010100101010
10011000010101 0101010

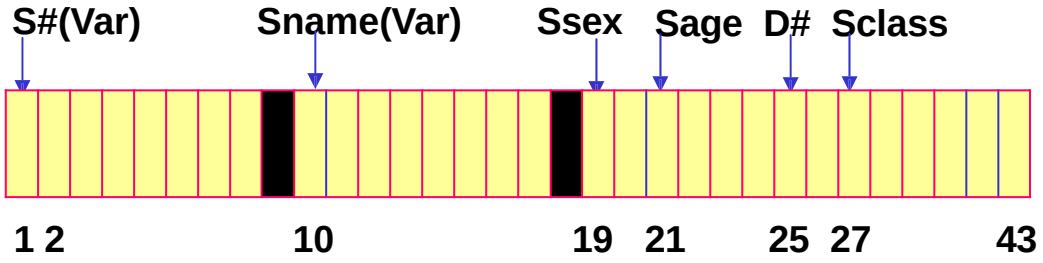
4. 数据库之表-记录与磁盘块的映射

(2)数据库中记录的区分及记录内属性值的区分

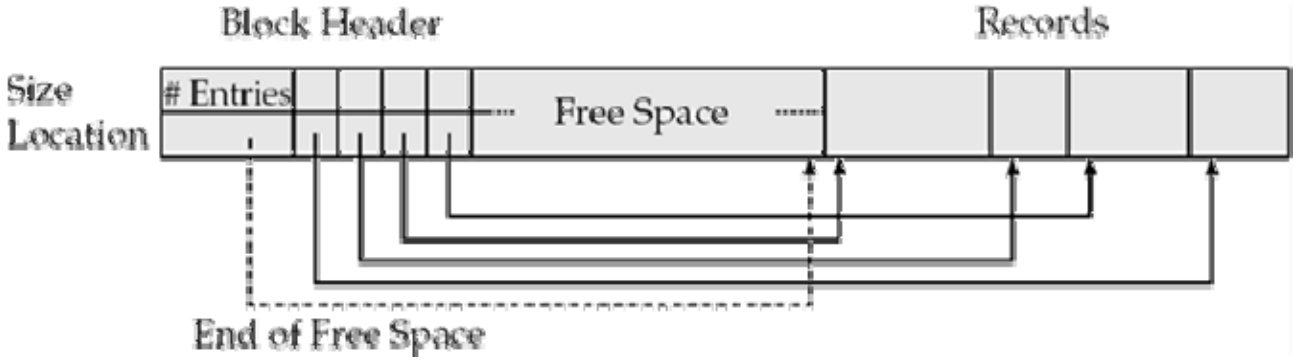
数据库记录在磁盘上的存储

❒定长记录，还是变长记录(靠分隔符区分开始与结束)





- 按长度区分记录
- 按指针(或标志)区分记录
- 块头如何设计



4. 数据库之表-记录与磁盘块的映射

(3)数据库中的记录 vs. 磁盘块

- 记录是非跨块存储，还是跨块存储(靠指针连接)

4. 数据库之表-记录与磁盘块的映射

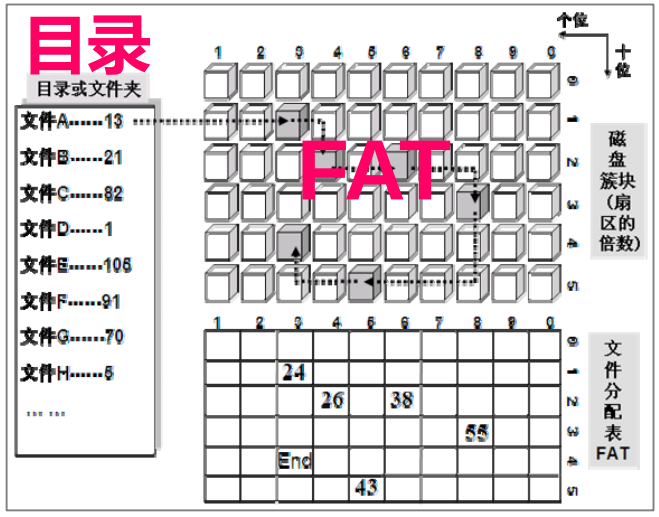
(4)数据库的表 vs. 磁盘块

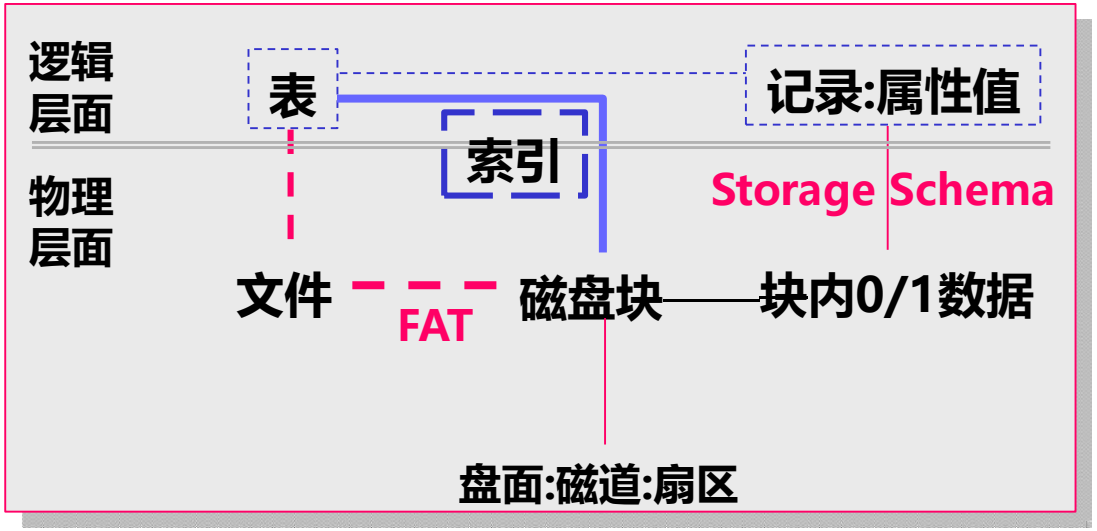
❑数据库-表所占磁盘块的分配方法

- ✓ 连续分配: 数据块被分配到连续的磁盘块上(会存在扩展困难问题)
- ✓ 链接分配: 数据块中包含指向下一数据块的指针(访问速度问题)
- ✓ 按簇分配: 按簇分配，簇之间靠指针连接; 簇有时也称片段

Segment或盘区extent

- ✓ 索引分配: 索引块中存放指向实际数据块的指针





第17讲 数据库物理存储

1. 基础回顾-计算机系统的存储体系
2. 磁盘的结构与特性
3. DBMS数据存储与查询实现的基本思想
4. 数据库之表和记录与磁盘块的映射
5. 数据库之文件组织方法?

5. 数据库之文件组织方法

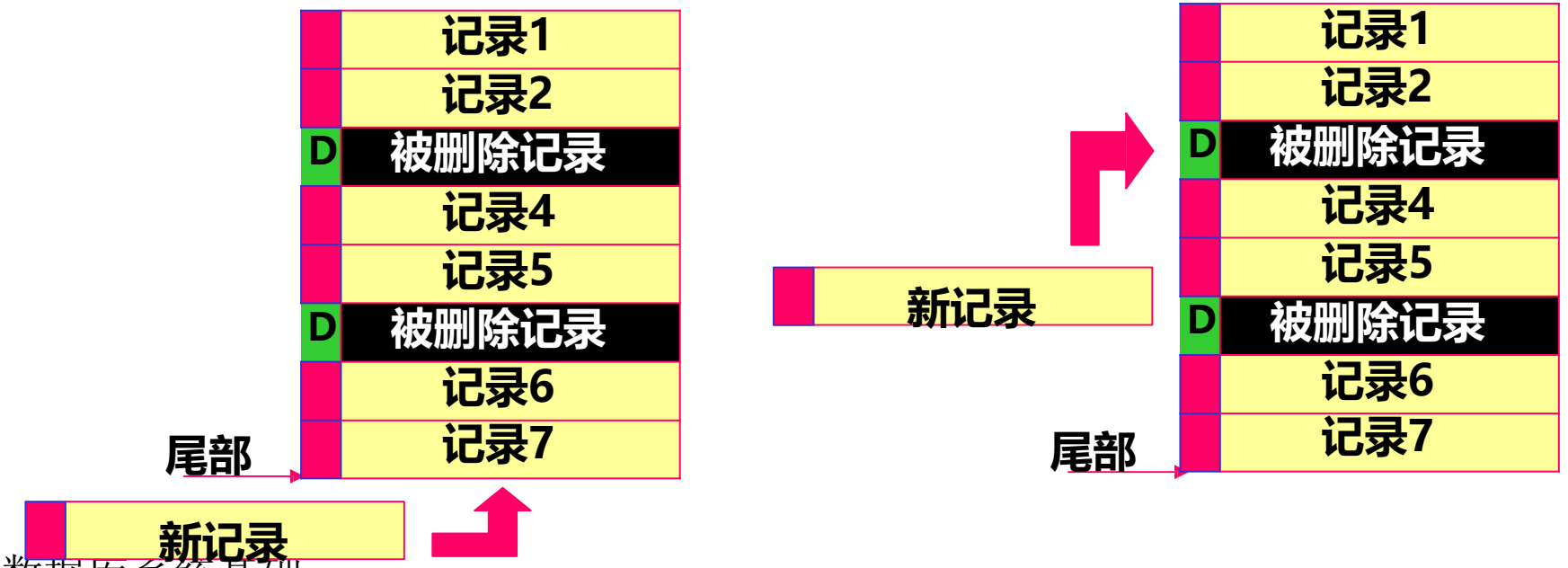
(1)数据组织与存取方法

- **数据组织要考虑更新(增、删、改)和检索需求**
  - ✓ 更新将涉及数据存储空间的扩展与回收问题
  - ✓ 检索将涉及扫描整个数据库的问题、大批量处理数据问题
  - ✓ 不同的需求要求不同的数据组织方法和存取方法
  - ✓ **文件组织(File Organization)**指的是数据组织成记录、块和访问结构的方式，包括把记录和块存储在磁盘上的方式，以及记录和块之间相互联系的方法
  - ✓ **存取方法(Access Method)**指的是对文件所采取的存取操作方法
  - ✓ 一种文件组织可以采取多种存取方法进行访问

5. 数据库之文件组织方法

(2)无序文件组织

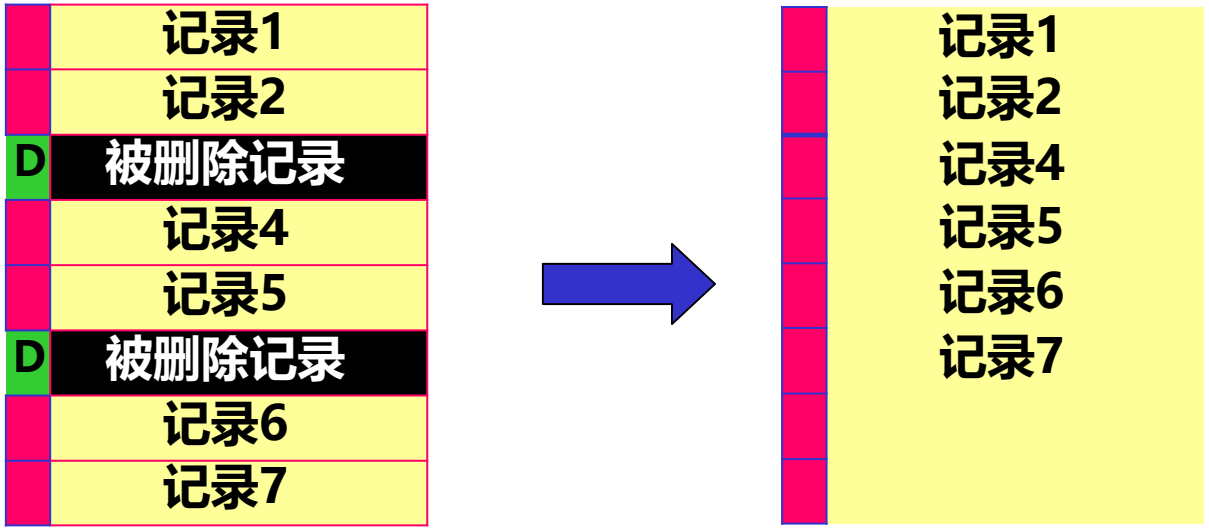
- **文件组织方法之一：无序记录文件(堆文件heap或pile file)**
  - ✓ 特点：记录可存储于任意有空间的位置，磁盘上存储的记录是无序 的。更新效率高，但检索效率可能低
  - ✓ 方法1：新记录总插入到文件尾部；删除记录时，可以直接删除该记录 所在位置的内容，也可以在该记录前标记“删除标记”
  - ✓ 方法2：在前者基础上，新增记录可以利用那些标记为“删除标记”的 记录空间



5. 数据库之文件组织方法

(2)无序文件组织

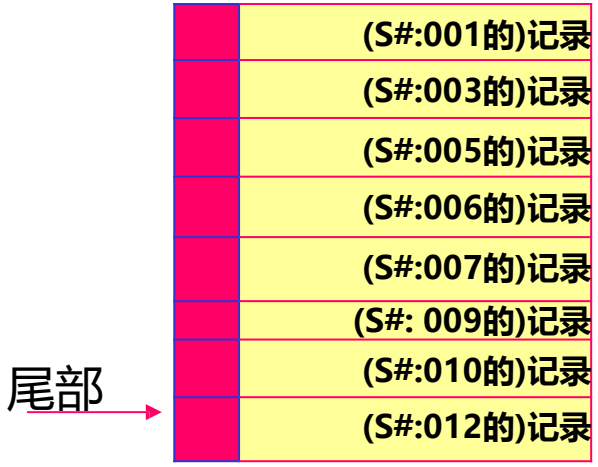
- **文件组织方法之一：无序记录文件(堆文件heap或pile file)(续)**
  - ✓ 频繁删增记录时会造成空间浪费，所以需要周期性重新组织数据库
  - ✓ **数据库重组(Reorganization)**是通过移走被删除的记录使有效记录连续存放，从而回收那些由删除记录而产生的未利用空间。



5. 数据库之文件组织方法

(3)有序文件组织

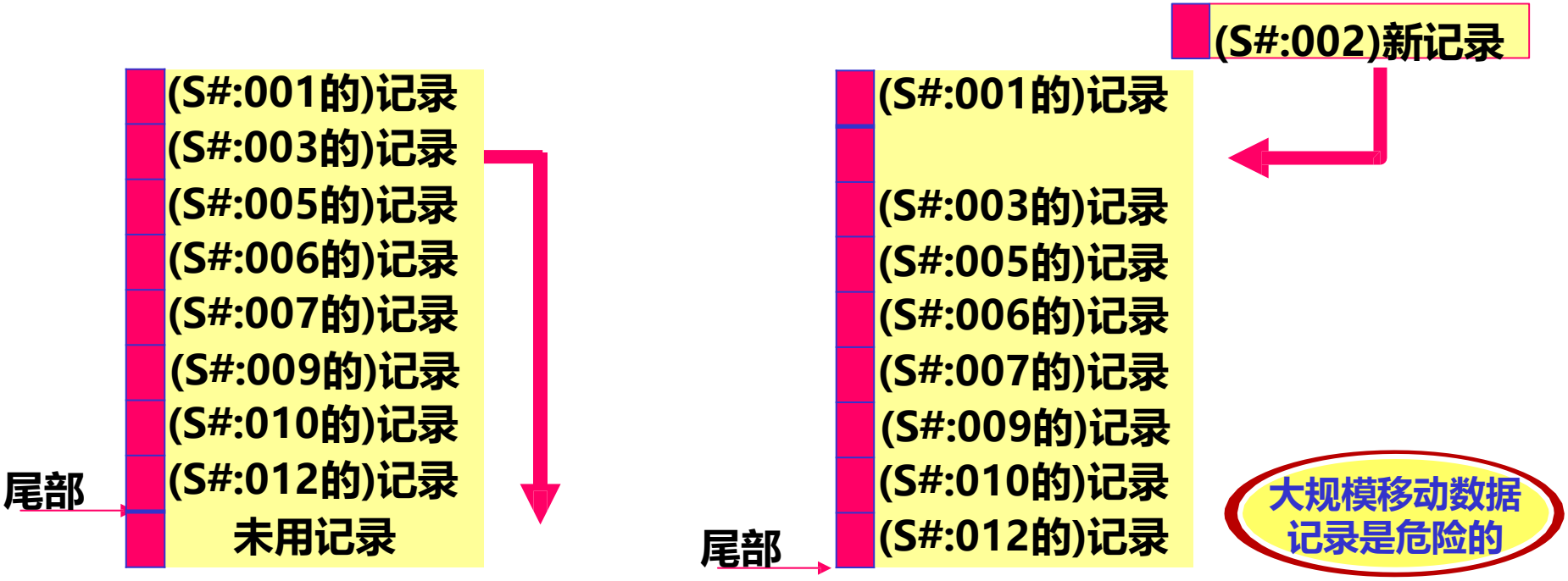
- **文件组织方法之二：有序记录文件(排序文件Sequential)**
  - ✓ 特点：记录按某属性或属性组值的顺序插入，磁盘上存储的记录是有序的。检索效率可能高。
  - ✓ 用于存储排序的属性通常称为排序字段(Ordering field)，通常，排序字段使用关系中的主码， 所以又称排序码(Ordering key)
  - ✓ 当按排序字段进行检索时，速度得到很大提高；但当按非排序字段检索时，速度可能不会提高很多



5. 数据库之文件组织方法

(3)有序文件组织

- **文件组织方法之二：有序记录文件(排序文件Sequential)(续)**
  - ✓ 有序记录文件的更新效率可能很低
  - ✓ 因为：在更新时要移动其他记录，为插入记录留出空间



5. 数据库之文件组织方法

(3)有序文件组织

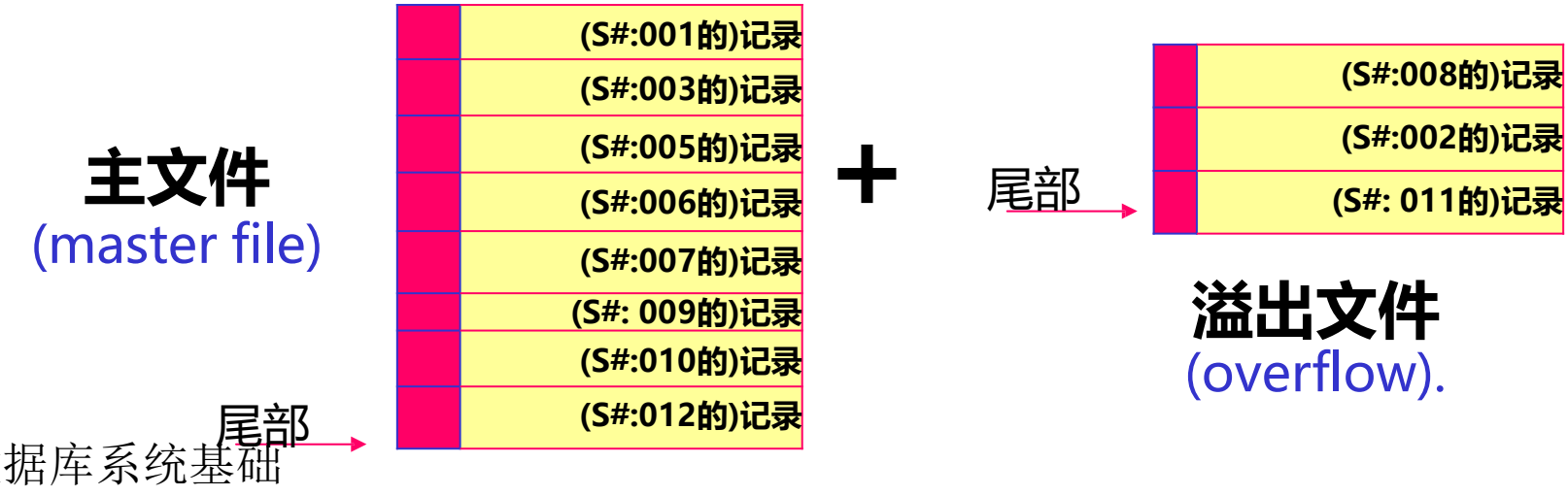
❑ **文件组织方法之二：有序记录文件(排序文件Sequential) (续)**

✓ 改进办法（使用溢出）：为将来可能插入元组预留空间(这可能造成空间浪费), 或使用一个临时的无序文件(被称为溢出文件)保留新增的记录。

✓当采取溢出文件措施时，检索操作既要操作主文件，又要操作溢出文件。

✓所以需要周期性重新组织数据库

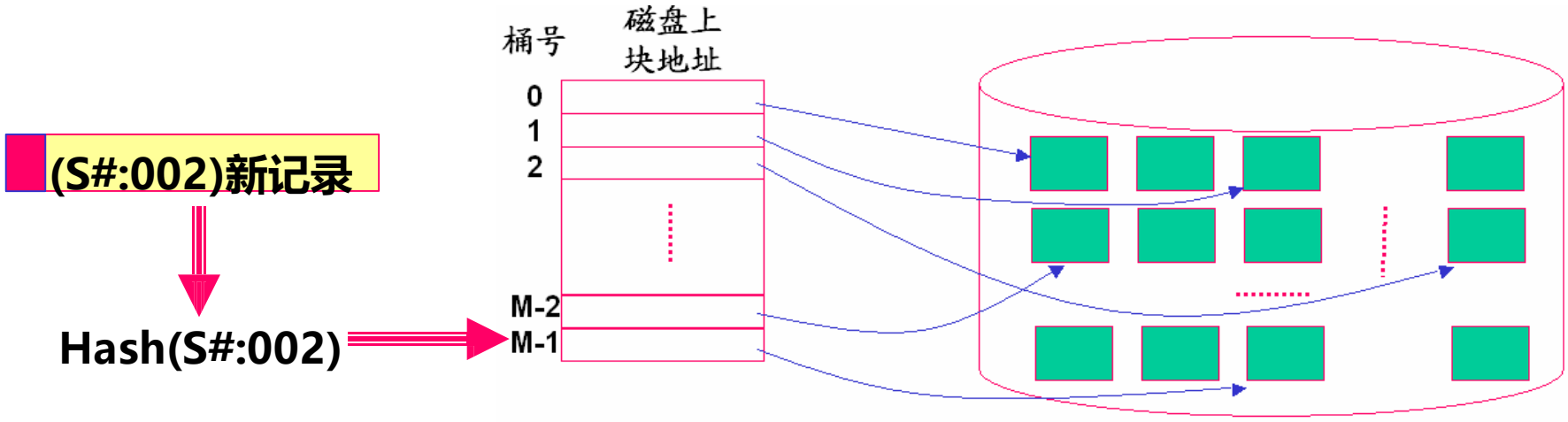
✓**数据库重组**将溢出文件合并到主文件，并恢复主文件中的记录顺序。



5. 数据库之文件组织方法

(4)散列文件组织

- **文件组织方法之三：散列文件(Hash file)**
  - ✓ 特点：可以把记录按某属性或属性组的值，依据一个散列函数来计算其应存放的位置：桶号(Bucket，块号或簇号等)。检索效率和更新效率都有一定程度的提高
  - ✓ 用于进行散列函数计算的属性通常称为散列字段(Hash field)，散列字段通常也采用关系中的主码，所以又称散列码(hash key)



5. 数据库之文件组织方法

(4)散列文件组织

- **文件组织方法之三：散列文件(Hash file)(续)**
  - ✓ 不同记录可能被hash成同一桶号，此时需在桶内顺序检索出某一记录

bucket 0

bucket 1

bucket 2

bucket 3

A-217	Brighton	750
A-305	Round Hill	350

bucket 4

A-222	Redwood	700

bucket 5

A-102	Perryridge	400
A-201	Perryridge	900
A-218	Perryridge	700

bucket 6

bucket 7

A-215	Mianus	700

bucket 8

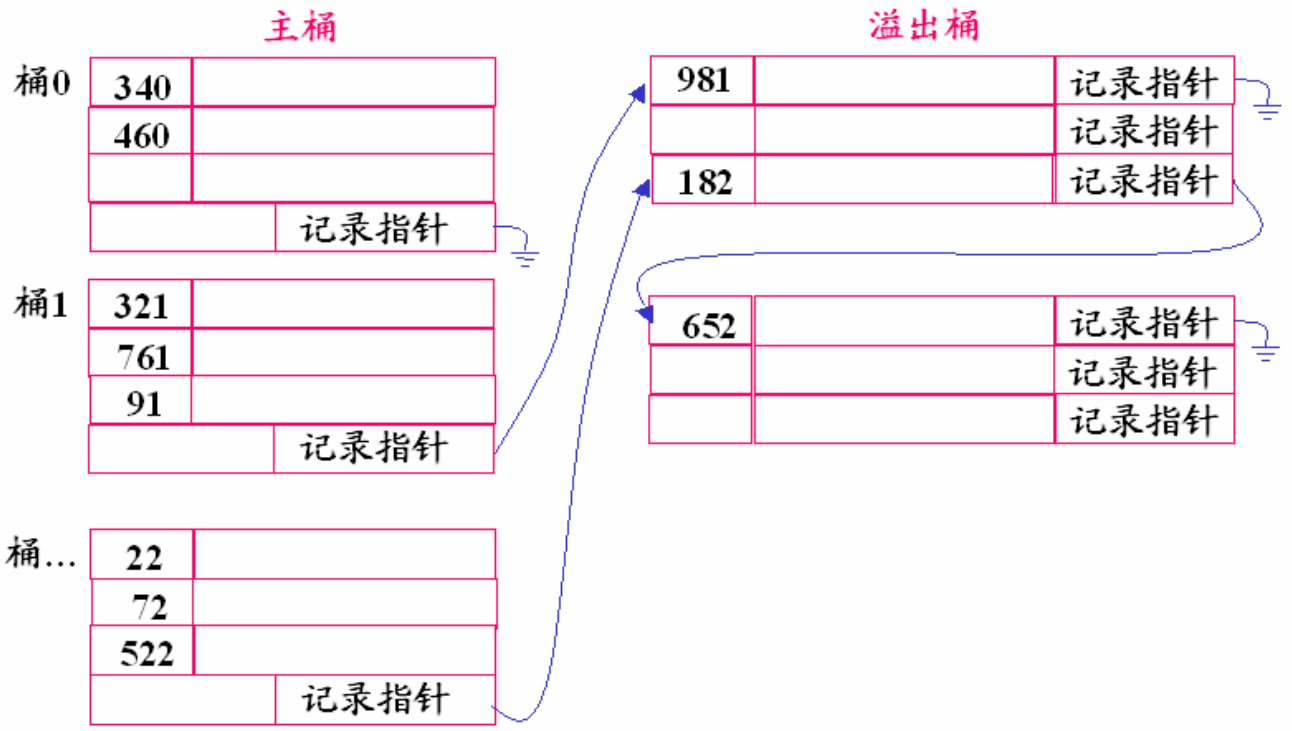
A-101	Downtown	500
A-110	Downtown	600

bucket 9

5. 数据库之文件组织方法

(4)散列文件组织

- **文件组织方法之三：散列文件(Hash file)(续)**
  - ✓ 链接法处理溢出
  - ✓ 散列还有许多问题及许多的处理技巧, 如散列桶的数目以及桶的大小, 动态散列技术等等(这里不再叙述)



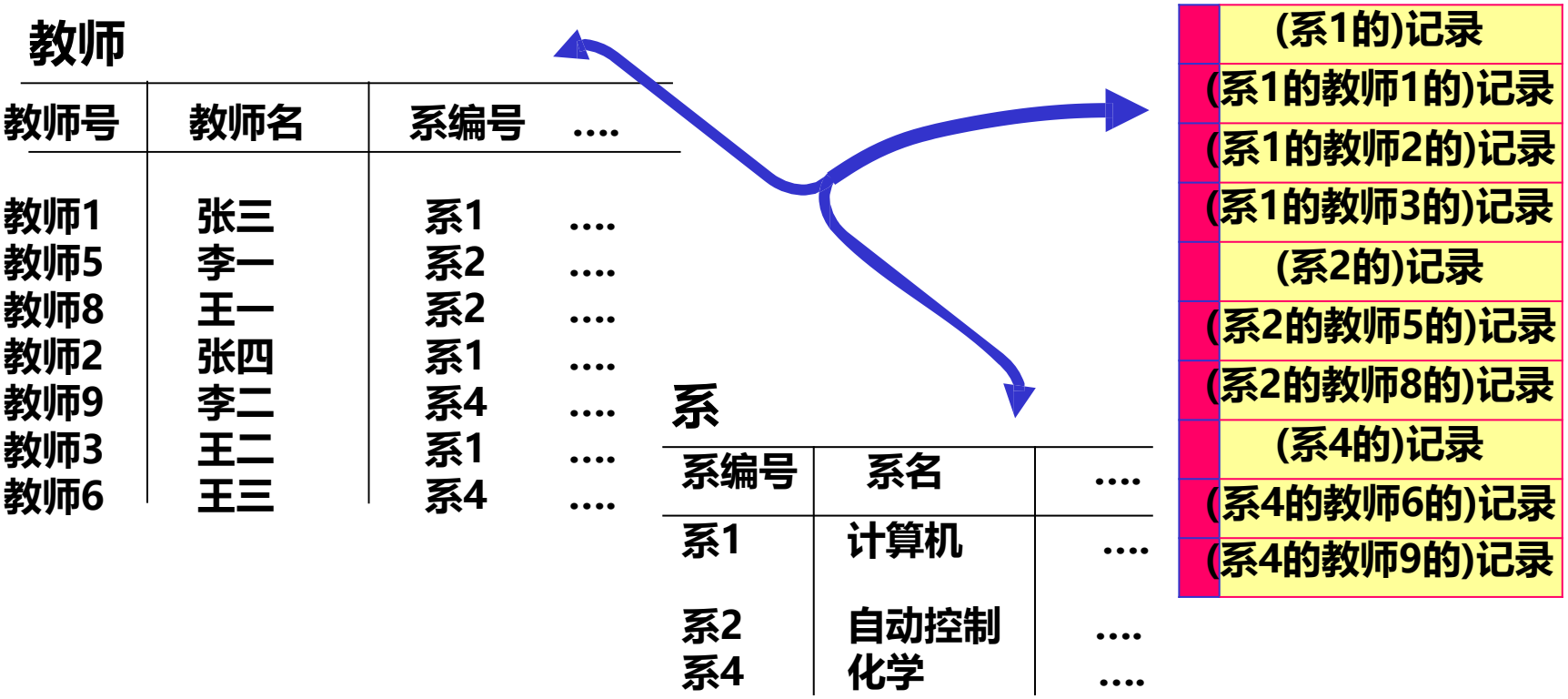
5. 数据库之文件组织方法

(4)聚簇文件组织

❑ 文件组织方法之四：聚簇文件(Clustering file)

✓聚簇：将具有相同或相似属性值的记录存放于连续的磁盘簇块中

✓多表聚簇：将若干个相互关联的Table存储于一个文件中—这可提高多表情况下的查询速度(有很多问题及相关的处理技巧，不再详细叙述)



5. 数据库之文件组织方法

(5)小结

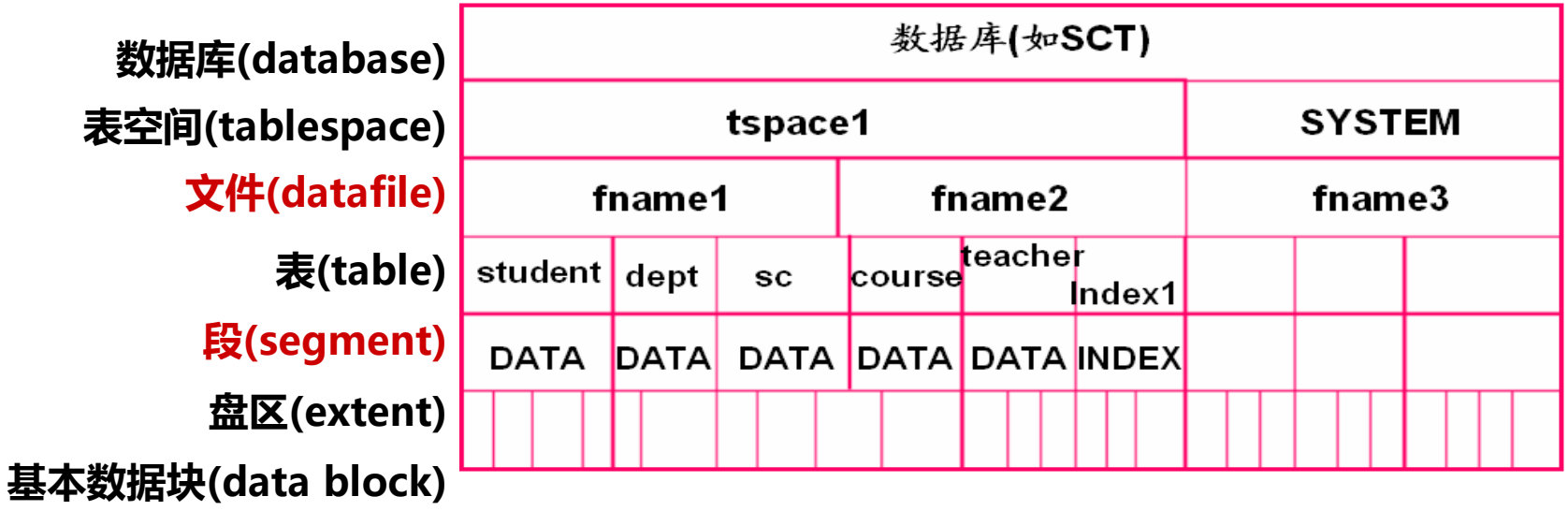
增删改时如何快速“存”

检索查询时如何快速“取”

Oracle DB物理存储简介

(1)Oracle数据库的数据组织

➢**Oracle数据库组织为数据库(database)、表空间(tablespace)、操作系统文件、table、段(segment)、盘区(extent)和基本数据块(data blocks)**



Oracle DB物理存储简介

(1)Oracle数据库的数据组织

➢**简要理解**

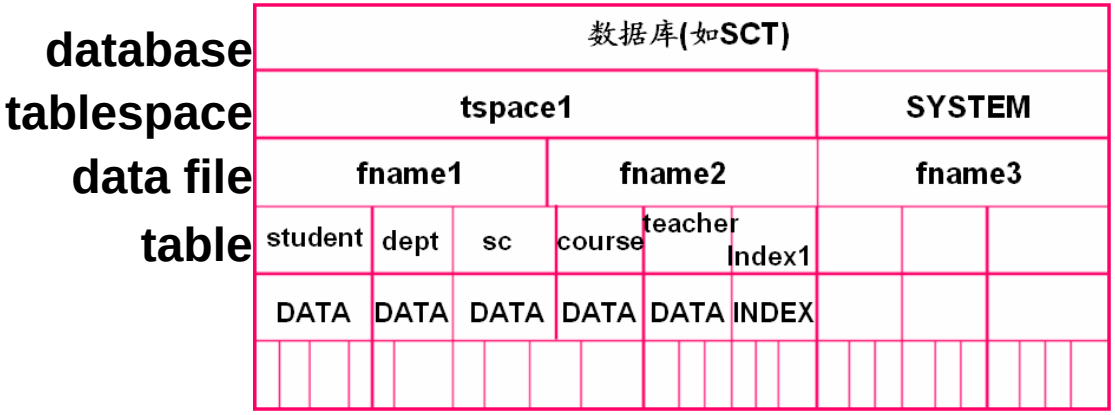
●每个数据库分成一个或多个表空间，所有表空间的组合存储容量即是数据库的存储容量

●有系统表空间SYSTEM和用户表空间；系统表空间由Oracle在创建数据库时自动创建，用于数据字典等的管理；用户表空间可由用户创建

●每个表空间由一个或多个操作系统文件构成，一个操作系统文件只能与一个数据库相联，操作系统文件仅起占位的作用。数据在表空间中可跨文件进行操作。

●操作系统文件中存储一个表(Table)或多个表，一个表可存储在一个文件中也可能存储在多个文件中。

●上述为逻辑存储层

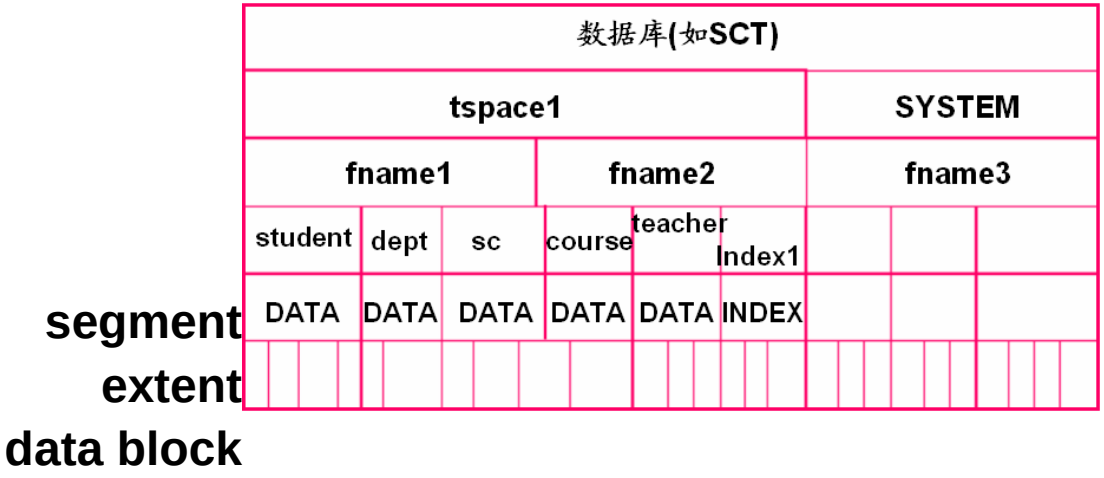


Oracle DB物理存储简介

(1)Oracle数据库的数据组织

➢简要理解(续)

- 物理存储层由段(segment)、盘区(extent)和数据块(data block)构成
- 数据块是最小的IO存储单位，又称为Oracle块、页(相当于扇区)
- 盘区是特定数量的连续数据块(相当于簇)。Oracle中的盘区是可以动态变化的，随不同数据库存储需求而调整
- 段是一组分配了特定数据结构的盘区，又分为数据段、索引段和临时段等
- 一个表的数据可以存放在一个段内，也可以存放在多个段内。一个段可以存放一个表的数据，也可以存放多个表的数据(如聚簇文件)



Oracle DB物理存储简介

(2)Oracle数据库的数据组织相关的SQL命令

- **Oracle物理数据库相关的定义语句**
- **定义表空间时，需要考虑物理存储性能。注意参数的使用，参数的含义参见Oracle手册。**

```
CREATE   TABLESPACE    tblspname
        DATAFILE   'filename'
                [SIZE n  [K|M]]
                [REUSE]
                [AUTOEXTEND OFF | AUTOEXTEND ON
                      [NEXT n [K|M] [MAXSIZE {UNLIMITED | n [K|M]}]
                {, 'filename' . . .}
        [ONLINE | OFFLINE]
        [DEFAULT STORAGE ([ INITIAL n] [NEXT n] [MINEXTENTS n]
                [MAXEXTENTS {n | UNLIMITED}] [PCTINCREASE n])]
        [MINIMUM EXTENT n [K|M]] ;
```

Oracle DB物理存储简介

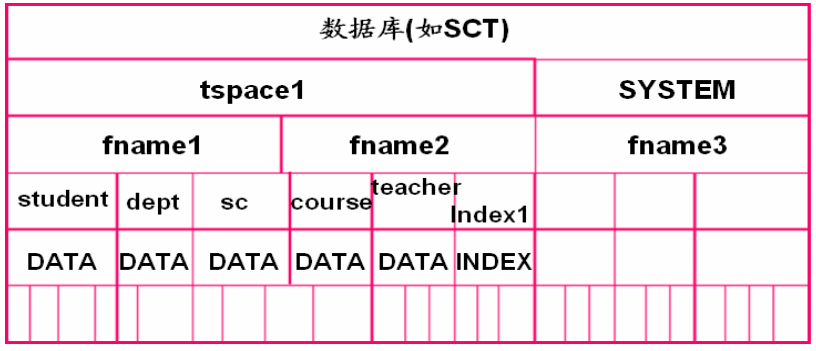
(2)Oracle数据库的数据组织相关的SQL命令

➢**简要理解**

一个表空间可以由多个文件构成

表空间的文件是否可自动扩展

```
CREATE  TABLESPACE   tblspname
        DATAFILE  'filename'
                [SIZE n  [K|M]]
                [REUSE]
                [AUTOEXTEND OFF | AUTOEXTEND ON
                    [NEXT n [K|M] [MAXSIZE {UNLIMITED | n [K|M]}]
                {, 'filename' . . .}
        [ONLINE | OFFLINE]
        [DEFAULT STORAGE ([ INITIAL n] [NEXT n] [MINEXTENTS n]
                [MAXEXTENTS {n | UNLIMITED}] [PCTINCREASE n])]
        [MINIMUM EXTENT n [K|M]] ;
```



可设置表空间的初始容量、最小盘区数、每次扩展容量等

盘区的大小是可以动态调整的

可设置构成表空间的最小盘区

Oracle DB物理存储简介

(2)Oracle数据库的数据组织相关的SQL命令

➢**定义表(table)时，需要考虑物理存储性能。注意参数的使用，参数的含义参见Oracle手册。**

```
CREATE   TABLE    tablename
        ( {colname datatype [default  { constant | NULL}]
                {, colname datatype etc. . . .}
        [ORGANIZATION HEAP | ORGANIZATION INDEX ]
        [TABLESPACE   tblspname]
        [STORAGE  ([initial n [K|M]]
                   [next n [K|M]]   [minextents n]
                   [maxextents n] [pctincrease n] ) ]
          [PCTFREE n]
        [PCTUSED n]
```

SQL的CREATE TABLE的三种功能：
(1)定义模式；(2)定义物理存储结构；(3)定义完整性约束

Oracle DB物理存储简介

(2)Oracle数据库的数据组织相关的SQL命令

```
CREATE  TABLE   tablename
        ( {colname datatype [default  { constant | NULL}]
                {, colname datatype etc. . . .}
        [ORGANIZATION HEAP | ORGANIZATION INDEX ]
        [TABLESPACE  tblspname]
        [STORAGE  ([initial n [K|M]]
                   [next n [K|M]]   [minextents n]
                   [maxextents n]   [pctincrease n] ) ]
        [PCTFREE n]
        [PCTUSED n]
```

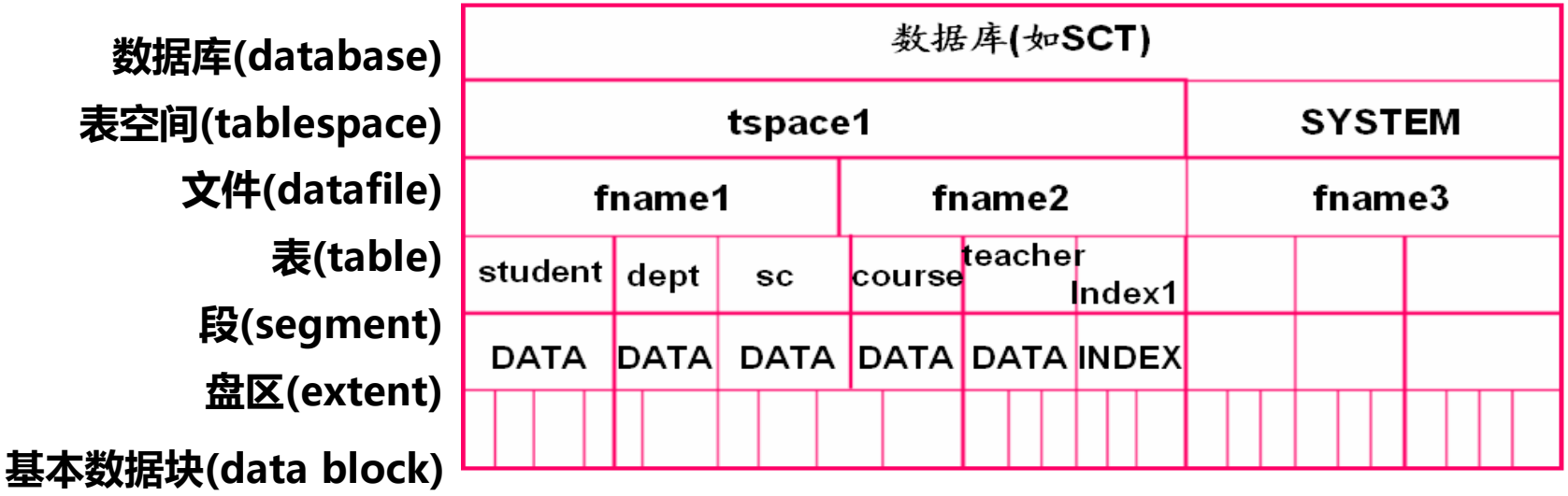
Initial # 表示分配给表（段）的初始盘区的大小
next # 指定第二个盘区的大小
pctincrease # 指定第三个及后续分配的盘区的增长百分比
minextents # 指定创建表时，至少要分配多少个盘区给这个表（段）
maxextents # 指定可以给这个表（段）的盘区的最大数量

Pctused # 制定块中数据使用空间的最低百分比。
pctfree # 为存储块中行的更新预留空闲空间的最小百分比。

例如，假定在 Create table 语句中指定了 pctfree 为 20，则说明在该表的数据段内每个数据块的 20%被作为可利用的空闲空间，用于更新已在数据块内存在的数据行。其余 80%是用于插入新的数据行，直到达到 80%为止。显然，pctfree 值越小，则为现存行更新所预留的空间越少。因此，如果 pctfree 设置得太高，则在全表扫描期间增加 I/O，浪费磁盘空间；如果 pctfree 设置得太低，则会导致行迁移。

Oracle DB物理存储简介

(3)小结



CREATE DATABASE sct

将多个表空间关联在一起

CREATE TABLESPACE tspace1…

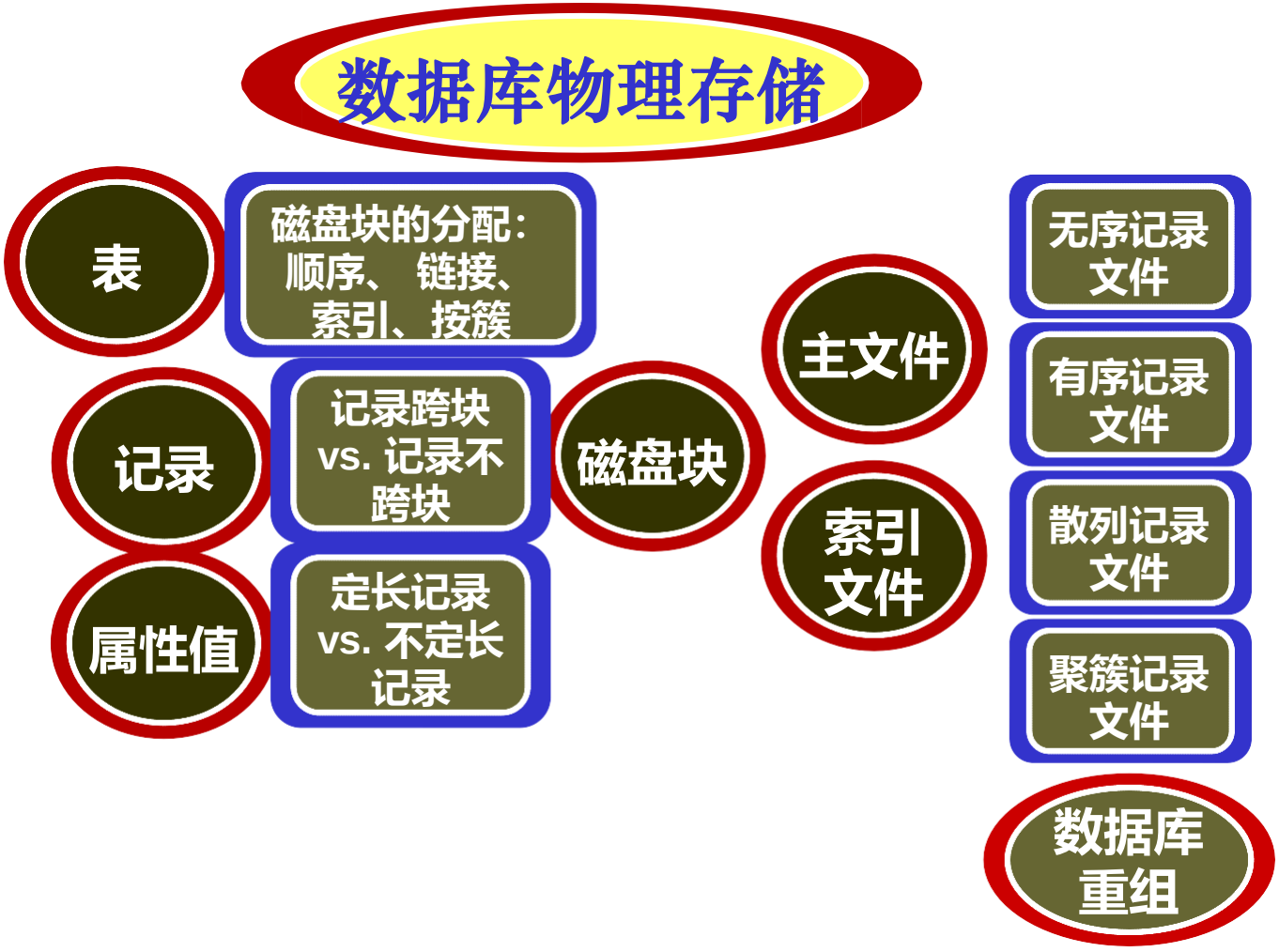
用多个操作系统文件占位，将多个表的存取统一在一个空间中

直接对数据块-盘区-Segment进行操作。

CREATE TABLE tablename … 表的存取，直接映射对数据块-盘区的操作

回顾本讲学习了什么?

回顾本讲学习了什么?


数据库物理存储
表
磁盘块的分配：顺序、 链接、索引、按簇
记录
记录跨块 vs. 记录不跨块
磁盘块
属性值
定长记录 vs. 不定长记录
主文件
索引文件
无序记录文件
有序记录文件
散列记录文件
聚簇记录文件
数据库重组