Capítulo 11: Almacenamiento y estructura de archivos

Fundamentos de Bases de datos, 5ª Edición.

Capítulo 11: Almacenamiento y estructura de archivos

- Visión general de los medios físicos de almacenamiento
- Discos magnéticos
- RAID
- Almacenamiento terciario
- Acceso al almacenamiento
- Organización de archivos
- Organización de los registros en archivos
- Almacenamiento del diccionarios de datos
- Estructuras de almacenamiento para las bases de datos orientadas a objetos

Clasificación de los medios físicos de almacenamiento

- Velocidad a la que se puede acceder a los datos
- Coste por unidad de datos
- Fiabilidad
  - pérdida de datos por fallos del suministro eléctrico o caídas del sistema
  - fallos físicos de los dispositivos de almacenamiento
- El almacenamiento se puede dividir en:
  - **almacenamiento volátil:** el contenido se pierde cuando se corta el suministro eléctrico
  - **almacenamiento no volátil**:
    - El contenido se conserva cuando se corta el suministro eléctrico.
    - Incluye almacenamiento secundario y terciario, así como memoria principal con batería de salvaguarda.

Medios físicos de almacenamiento

- **Cache** – la forma más rápida y costosa de almacenamiento; volátil; gestionada por el hardware del sistema informático. No es gestionada por los motores de base de datos.
- **Memoria principal**:
  - acceso rápido (de 10 a 100 nanosegundos; 1 nanosegundo = 10^{-9} segundos)
  - generalmente demasiado pequeña (o demasiado cara) para almacenar toda la base de datos
    - capacidades de hasta unos pocos Gigabytes, ampliamente usada en la actualidad
    - La capacidades han aumentado y los costes por byte han disminuido de forma constante y rápida (aproximadamente un factor 2 cada 2 ó 3 años)
  - **Volátil** — los contenidos de la memoria principal generalmente se pierden si se produce un fallo en el suministro eléctrico o una caída del sistema.

Medios físicos de almacenamiento (cont.)

- **Memoria flash**
  - Los datos superan los fallos del suministro eléctrico
  - Se pueden escribir datos en una posición sólo una vez, pero la posición puede ser borrada y grabada de nuevo
    - Sólo puede soportar un número limitado de ciclos (10K – 1M) de escritura / borrado.
    - El borrado de la memoria ha de hacerse sobre una banco entero de memoria
  - Las lecturas son aproximadamente tan rápidas como en memoria principal
  - Aunque las escrituras son lentas (unos pocos microsegundos), borrar es más lento
  - El coste por unidad de almacenamiento es aproximadamente igual al de la memoria principal
  - Usada ampliamente en dispositivos incorporados, tales como cámaras digitales
  - también conocido como EEPROM (Memoria de lectura y borrado programable eléctricamente)

Medios físicos de almacenamiento (cont.)

- **Disco magnético**
  - Los datos se almacenan sobre disco giratorios y son leídos / grabados magnéticamente
  - Soporte primario para el almacenamiento de datos a largo plazo; generalmente almacena toda la base de datos.
  - Para los accesos se deben mover los datos desde disco a memoria principal y para el almacenamiento se han de volver a escribir
    - Acceso mucho más lento que en memoria principal (más sobre ello posteriormente)
  - **acceso directo** – adecuado para las lecturas de datos en disco en cualquier orden, a diferencia de la cinta magnética
  - Rango de capacidades de hasta aproximadamente 700 GB en la actualidad
    - Mucha más capacidad y mejor coste por byte que en memoria principal / memoria flash
    - Creciendo constante y rápidamente, con perfeccionamiento tecnológico (un factor de 2 a 3 cada 2 años)
  - Sobrevive a los fallos de suministro eléctrico y las caídas del sistema
    - un fallo de disco puede destruir datos, pero es muy raro

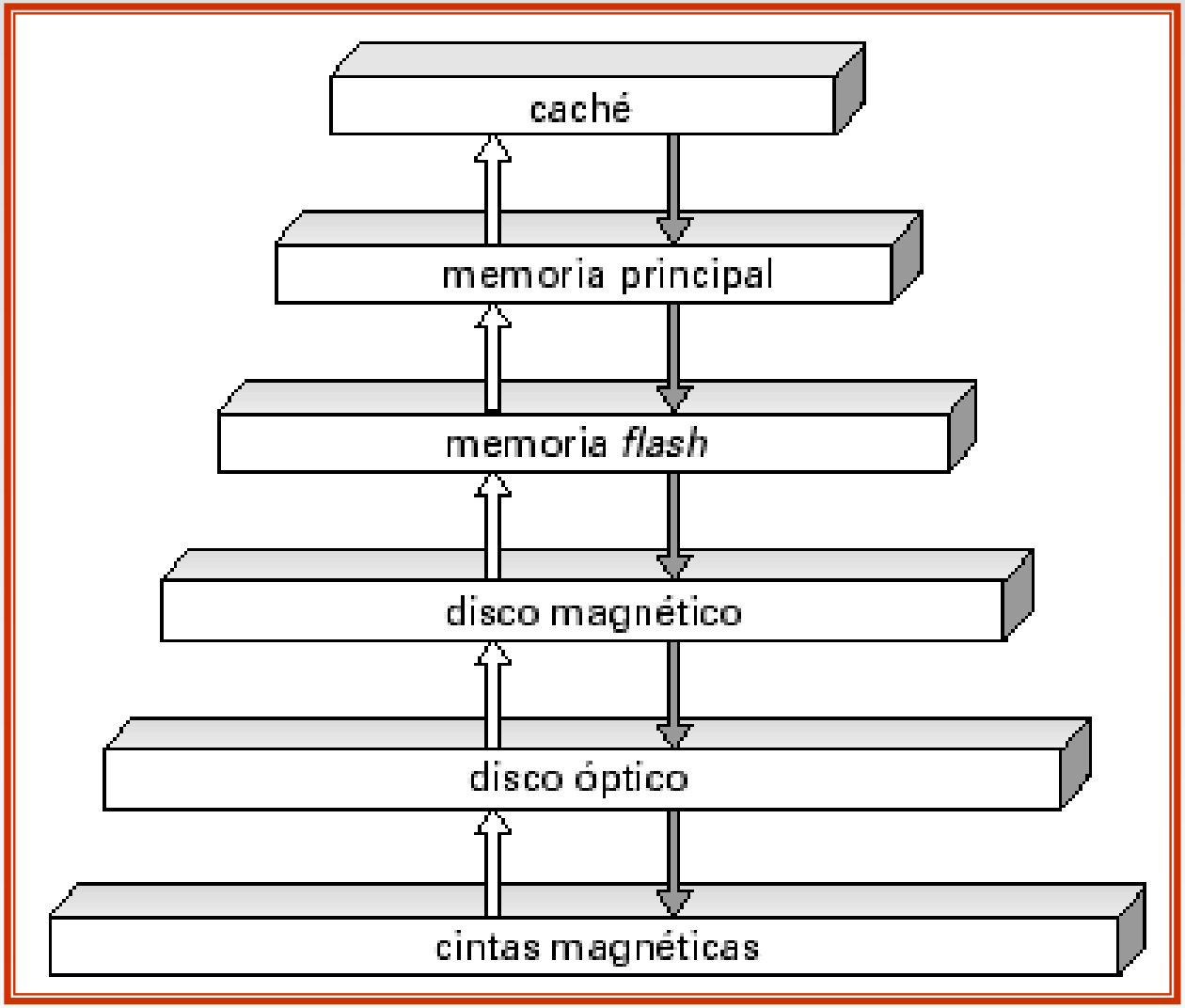
Medios físicos de almacenamiento (cont.)

- **Almacenamiento óptico**
  - no volátil, los datos se leen ópticamente, por medio de un láser, desde un disco giratorio
  - los formatos más populares son CD-ROM (640 MB) y DVD (4.7 hasta 17 GB)
  - Discos ópticos de escritura única y lectura múltiple, empleados para el almacenamiento de archivos (CD-R, DVD-R y DVD+R)
  - También están disponibles versiones de escritura múltiple (CD-RW, DVD-RW, DVD+RW y DVD-RAM)
  - Las lecturas y escrituras son más lentas que con discos magnéticos
  - Sistemas de Juke-box, con gran número de discos removibles, unos cuantos lectores y un mecanismo para la carga/descarga automática de discos, para el almacenamiento de grandes volúmenes de datos

Medios físicos de almacenamiento (cont.)

- **Almacenamiento en cinta**
  - no volátil, empleado principalmente para copias de seguridad (para la recuperación de un fallo de disco) y para datos de archivo
  - **acceso secuencial** – mucho más lento que los discos
  - capacidad muy alta (cintas disponibles de 40 a 300 GB)
  - El coste de almacenamiento es más barato que el disco pero las unidades de cinta son caras
  - Cambiadores de cintas disponible para el almacenamiento de cantidades masivas de datos
    - desde cientos de terabytes (1 terabyte = 10^9 bytes) hasta incluso un petabyte (1 petabyte = 10^{12} bytes)

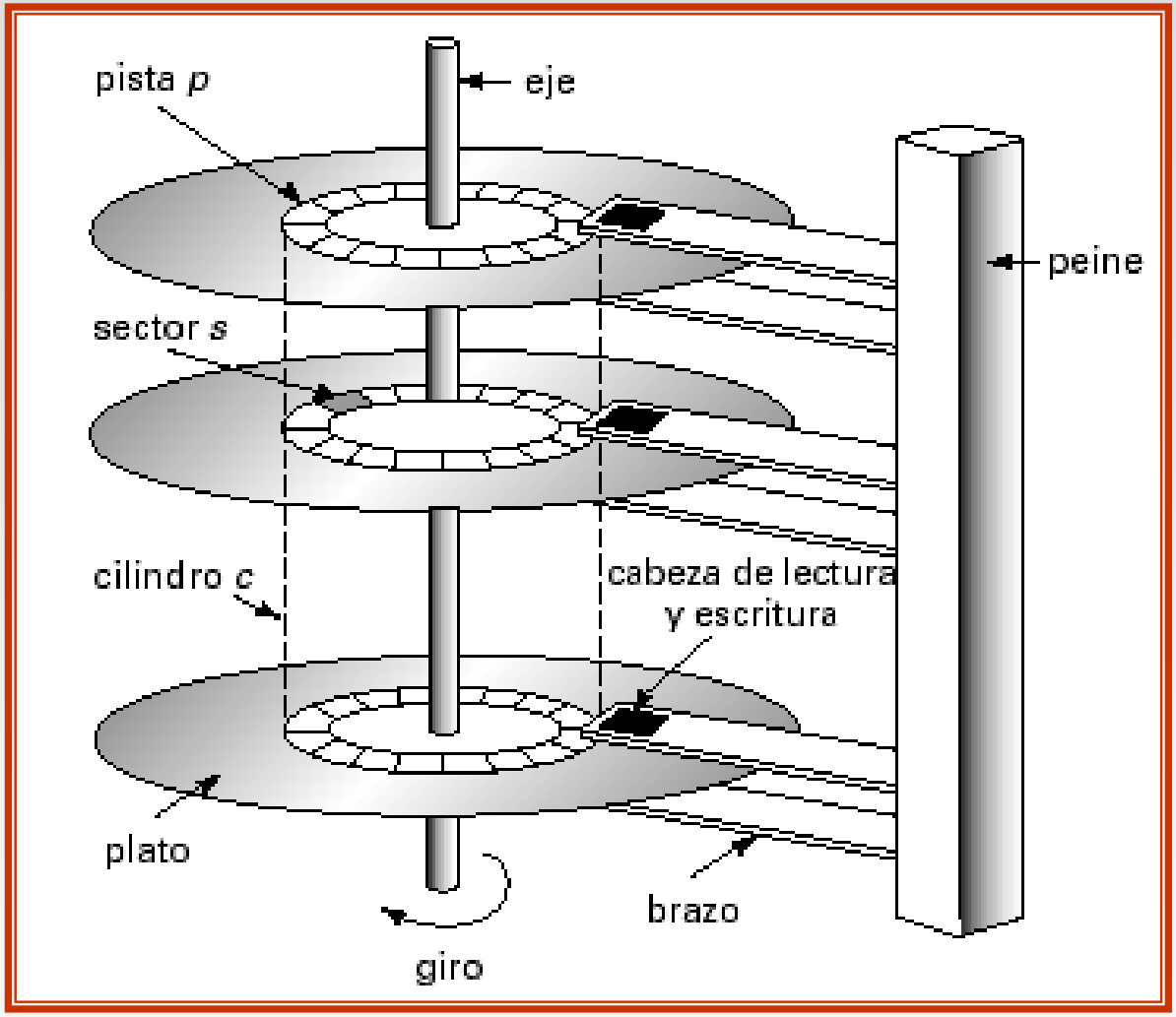
Jerarquía de almacenamiento



Jerarquía de almacenamiento (cont.)

- **almacenamiento principal**: Es el medio más rápido, pero volátil (caché, memoria principal).
- **almacenamiento secundario**: siguiente nivel jerárquico, no volátil, tiempos de acceso moderadamente rápidos
  - también llamado almacenamiento en conexión
  - Por ejemplo, memoria flash, discos magnéticos
- **almacenamiento terciario**: nivel jerárquico más bajo, no volátil, tiempo de acceso lento
  - también llamado almacenamiento sin conexión
  - Por ejemplo, cintas magnéticas, almacenamiento óptico

Mecanismos de discos rígidos magnéticos



NOTA: El diagrama es esquemático y simplifica la estructura de los controles de discos reales

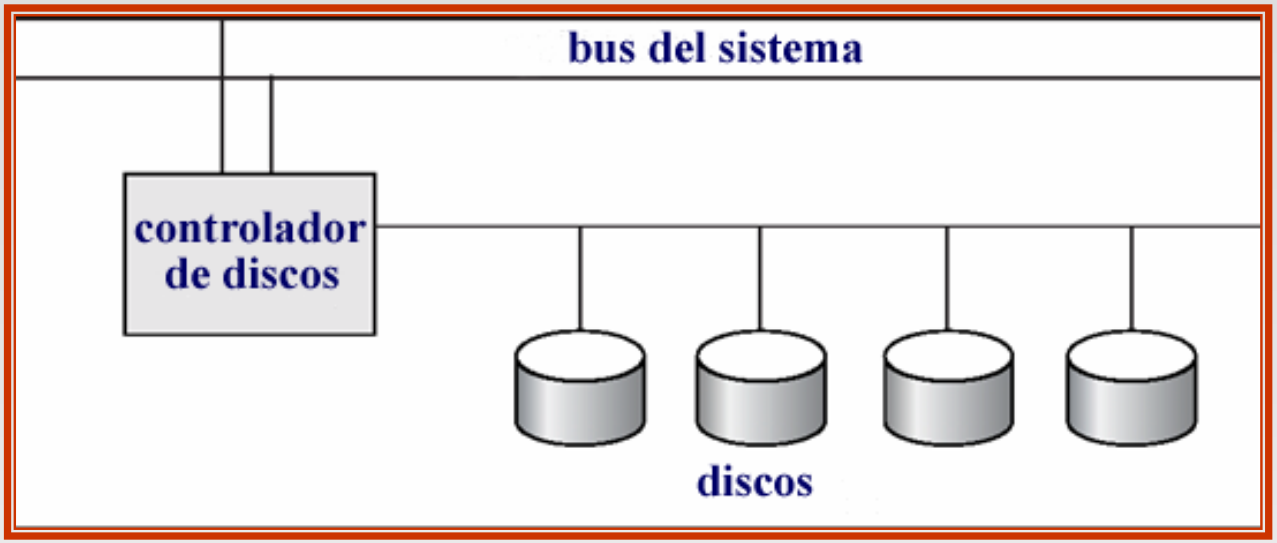
Discos magnéticos

- **Cabeza de lectura y escritura**
  - Colocado muy cerca de la superficie del plato (casi tocándolo)
  - Lee o escribe magnéticamente información codificada.
- La superficie del plato está dividida en **pistas** circulares
  - Alrededor de 50K-100Kpistas por plato en los discos rígidos típicos
- Cada pista está dividida en **sectores.**
  - Un sector es la unidad más pequeña de datos que puede ser leída o escrita.
  - El tamaño típico del sector es de 512 bytes
  - Sectores típicos por pista: de 500 (en pistas internas) hasta 1000 (en pistas externas)
- Para leer / escribir un sector
  - el brazo del disco gira hasta situar la cabeza sobre la pista correcta
  - el plato gira continuamente; los datos se leen / escriben según el sector pasa bajo la cabeza
- Dispositivos cabeza-disco
  - múltiples platos de discos en un solo huso (generalmente de 1 a 5)
  - una cabeza por plato, montadas sobre un brazo común.
- **Cilindro** *i* consta de i^{th} pistas de todos los platos

Discos magnéticos (cont.)

- Los discos de anteriores generaciones estaban expuestos a las caídas de las cabezas
  - Las superficies de los discos de anteriores generaciones tenían óxidos metálicos que se descomponían sobre la cabeza caída y dañaban todos los datos del disco
  - Los discos de las generaciones actuales están menos expuestos a tales desastres, aunque se pueden corromper sectores individuales
- **Controlador de disco** – interfaces entre el sistema informático y el hardware de la unidad de disco.
  - acepta comandos de alto nivel para leer o escribir un sector
  - inicia acciones como mover el brazo del disco a la pista correcta y leer o escribir los datos
  - Calcula y añade **comprobaciones de suma** a cada sector, para verificar que los datos se vuelvan a leer correctamente
    - Si los datos está corruptos, muy probablemente la comprobación de suma almacenada no se corresponderá con la comprobación de suma recalculada
  - Asegura una escritura satisfactoria, volviendo a leer el sector después de escribirlo
  - Realiza reasignaciones de sectores defectuosos

Subsistema de discos



- Múltiples discos conectados a un sistema informático por medio de un controlador
  - La funcionalidad de los controladores (comprobación de suma, reasignación de sectores defectuosos) frecuentemente es llevada a cabo mediante discos individuales; reduce la carga sobre el controlador
- Familias de estándares de interfaz de discos
  - ATA (adaptador AT)
  - SATA (Serial ATA)
  - SCSI (interconexión de pequeños sistemas informáticos)
  - Diversas variantes de cada estándar (diferentes velocidades y capacidades)

Medidas del rendimiento de discos

- **Tiempo de acceso** – el tiempo que lleva desde que se solicita una lectura o escritura, hasta que comienza la transferencia de los datos. Consta de:
  - **Tiempo de búsqueda** – tiempo que se tarda en reposicionar el brazo sobre la pista correcta.
    - El tiempo medio de búsqueda es 1/2 del tiempo de búsqueda en el peor de los casos.
      - Sería 1/3 si todas las pistas tuvieran el mismo número de sectores y se ignorara el tiempo de arranque y parada del movimiento del brazo
    - de 4 a 10 milisegundos en discos típicos
  - **Latencia rotacional** – tiempo de acceso que le lleva al sector situarse debajo de la cabeza.
    - La latencia media es 1/2 de la latencia en el peor de los casos.
    - 4 a 11 milisegundos en discos típicos (5400 a 15000 r.p.m.)
- **Velocidad de transferencia de datos** – la velocidad a la se pueden recuperar los datos del disco o grabarse en él.
  - Velocidad máxima de 25 a 100 MB por segundo, menor en las pistas internas
  - Múltiples discos pueden compartir un controlador, por lo que también es importante poder gestionar la velocidad del controlador
    - Por ejemplo, ATA-5: 66 MB/s, SATA: 150 MB/s, SCSI-3: 40 MB/s
    - Canal de fibra (FC2Gb): 256 MB/s

Medidas de rendimiento (cont.)

- **Tiempo medio entre fallos (MTTF)** – el tiempo medio que se espera funcione el disco continuamente, sin ningún fallo.
  - Generalmente de 3 a 5 años
  - La probabilidad de fallo de los discos nuevos es muy pequeña, de acuerdo con un "MTTF teórico" de 500.000 a 1.200.000 horas para un disco nuevo
    - Por ejemplo, un MTTF de 1.200.000 horas para un disco nuevo significa que, dados 1.000 discos relativamente nuevos, en promedio fallará uno cada 1.200 horas
  - El MTTF disminuye con la edad de los discos

Optimización del acceso a los bloques del disco

- **Bloque** – una secuencia contigua de sectores de una sola pista
  - el dato se transfiere en bloques entre el disco y la memoria principal
  - rango de tamaños desde 512 bytes hasta varios kilobytes
    - Bloques más pequeños: más transferencias desde disco
    - Bloques más grandes: más espacio derrochado, debido a los bloques parcialmente llenos
    - El rango de tamaños de los bloques típicos hoy, va desde 4 hasta 16 kilobytes
- Los algoritmos de **planificación del brazo del disco** ordenan los accesos pendientes a las pistas, de manera que el movimiento del brazo del disco sea mínimo
  - **algoritmo del ascensor**: se mueve el brazo del disco en una dirección (desde las pistas exteriores a las interiores o viceversa), procesando la siguiente petición en esa dirección hasta que no haya más peticiones, en cuyo caso se invierte la dirección y se repite

Optimización del acceso a los bloques del disco (cont.)

- **Organización de archivos** – se optimiza el tiempo de acceso a los bloques, organizándolos para que se correspondan con la forma en que se accederá a los datos
  - Por ejemplo La información relacionada se almacena en el mismo cilindro, o en cilindros próximos.
  - Los archivos pueden **fragmentarse** a lo largo del tiempo
    - Por ejemplo, si los datos se insertan en / borran desde el archivo
    - O se distribuyen los bloques libres sobre el disco, y el archivo creado de nuevo tiene sus bloques distribuidos sobre el disco
    - El acceso secuencial a un archivo fragmentado origina un aumento del movimiento del brazo
  - Algunos sistemas tienen utilidades para defragmentar el sistema de archivos, a la hora de acelerar el acceso a los archivos

Optimización del acceso a los bloques del disco (cont.)

- Las **memorias intermedias de escritura no volátil** aceleran las escrituras en disco grabando los bloques directamente sobre una memoria intermedia RAM no volátil
  - RAM no volátil: RAM con batería de salvaguarda o memoria flash
    - Incluso si falla el suministro eléctrico, los datos están seguros y se grabarán sobre el disco cuando vuelva el suministro
  - El controlador graba sobre el disco siempre que el disco no tenga otras peticiones o hayan estado pendientes por algún tiempo
  - Las operaciones de la base de datos que requieren que los datos estén previamente almacenados en forma segura, pueden seguir adelante sin esperar a que se graben
  - *Se pueden reordenar las escrituras para minimizar el movimiento del brazo del disco*
- **Disco de registro histórico**– un disco dedicado al registro histórico secuencial de las actualizaciones sobre los bloques
  - Se usa exactamente como la RAM no volátil
    - Escribir sobre el disco de registro histórico es muy rápido, dado que no son necesarias búsquedas
    - No es necesario hardware especial (NV-RAM)
- Los sistemas de archivos generalmente reordenan las escrituras sobre el disco para mejorar el rendimiento
  - Los **sistemas de archivos diarios** graban datos en orden seguro sobre NV-RAM, o sobre el disco de registro histórico
  - Sin reordenación diaria: riesgo de corrupción de los datos del sistema de archivos

RAID

- **RAID: Arrays redundantes de discos independientes**
  - técnicas de organización de discos que gestionan un gran número de discos, aportando la visión de uno solo de
    - alta capacidad y alta velocidad mediante el uso de múltiples discos en paralelo y
    - alta fiabilidad por el almacenamiento redundante de datos, para que se puedan recuperar incluso si falla un disco
- La posibilidad de que falle algún disco de entre un conjunto de *N* discos, es mucho mayor que la posibilidad de que falle un determinado disco en solitario.
  - Por ejemplo, un sistema con 100 discos, cada uno con MTTF de 100.000 horas (aproximadamente 11 años), tendrá un MTTF del sistema de 1.000 horas (aproximadamente 41 días)
  - Técnicas del empleo de redundancia para evitar que la pérdida de datos sea crítica con gran número de discos
- Técnicas del empleo de redundancia para evitar que la pérdida de datos sea crítica con gran número de discos
  - Originariamente la I de RAID significaba “barato” (inexpensive)
  - Hoy se emplean los RAID por su alta fiabilidad y ancho de banda.
    - La “I” se interpreta como independiente

Mejora de la fiabilidad, vía redundancia

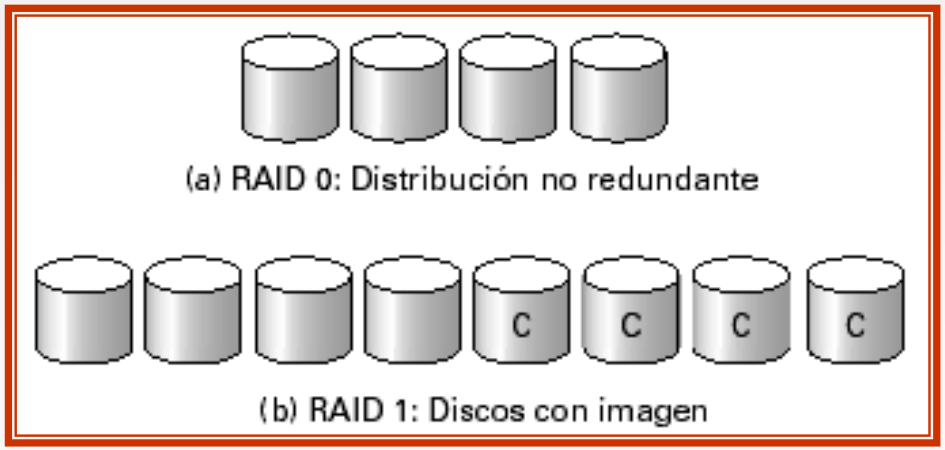
- **Redundancia** – información extra almacenada que se puede emplear para reconstruir la pérdida de información por el fallo de un disco
- Por ejemplo, **creación de imágenes** (o **creación de sombras**)
  - Duplicar cada disco. Un disco lógico consta de dos discos físicos.
  - Cada escritura se lleva a cabo en ambos discos
    - Las lecturas pueden tener lugar desde cualquiera de los discos
  - Si falla uno de los discos del par, los datos todavía están disponibles en el otro
    - La pérdida de datos sólo podría tener lugar si fallaran un disco y, antes de que se reparase el sistema, su disco imagen
      - La probabilidad de sucesos combinados es muy pequeña
        - Excepto por modos de fallos dependientes tales como un incendio, el hundimiento del edificio o una sobre tensión en el suministro eléctrico
- El tiempo medio entre pérdidas de datos depende del tiempo medio entre fallos, y del tiempo medio de reparación
  - Por ejemplo, MTTF de 100.000 horas, un tiempo medio de reparación de 10 horas da un tiempo medio entre pérdidas de datos de $500*10^6$ horas (ó 57.000 años) para un par de discos en imagen (ignorando modos de fallos dependientes)

Mejoras en el rendimiento vía paralelismo

- Dos objetivos principales del paralelismo en un sistema de discos:
  1. Equilibrar la carga en múltiples accesos pequeños, para incrementar la productividad
  2. Paralelizar los accesos grandes para reducir el tiempo de respuesta.
- Mejorar la velocidad de transferencia mediante la distribución de los datos a través de múltiples discos.
- **Distribución en el nivel de bit** – división de los bits de cada byte a través de múltiples discos
  - En un array de ocho discos, se escribe el bit *i* de cada byte sobre el disco *i.*
  - Cada acceso puede leer datos a ocho veces la velocidad de solo disco.
  - Sin embargo, el tiempo de búsqueda/acceso es peor que para un solo disco
    - La distribución en el nivel de bit no se usa mucho más
- **Distribución en el nivel de bloque** – con *n* discos, el bloque *i* de un archivo va al disco (*i* mod *n*) + 1
  - Se pueden ejecutar en paralelo peticiones para diferentes bloques, si los bloques residen en diferentes discos
  - Una petición para una secuencia grande de bloques puede utilizar todos los discos en paralelo

Niveles de RAID

- Esquemas para aportar redundancia al menor coste, empleando distribuciones de discos combinadas con bits de paridad
  - Diferentes organizaciones RAID, o niveles de RAID, tienen costes distintos, rendimientos y fiabilidades característicos
- **RAID de nivel 0: Distribución de bloques; sin redundancia.**
  - Se utiliza en aplicaciones de alto rendimiento en las que no es crítica la pérdida de datos.
- **RAID de nivel 1: Imágenes de discos** con distribución de bloques
  - Ofrece el mejor rendimiento de escritura.
  - Es popular en aplicaciones como el almacenamiento de archivos de registros históricos en un sistema de bases de datos.

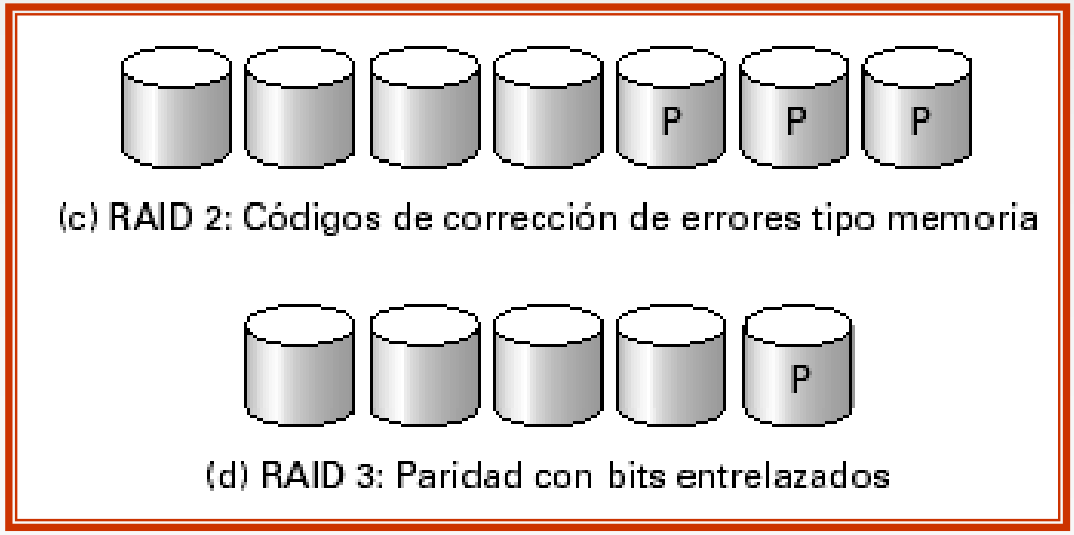


(a) RAID 0: Distribución no redundante

(b) RAID 1: Discos con imagen

Niveles de RAID (cont.)

- **RAID de nivel 2**: Organización de códigos de corrección de errores tipo memoria (ECC) con distribución de bit.
- **RAID de nivel 3**: Paridad con bits entrelazados
  - un solo bit de paridad es suficiente para la corrección de errores, no sólo detección, dado que se sabe el disco que ha fallado
    - Cuando se graban datos, los bits de paridad correspondientes se deben calcular y escribir sobre un disco de bit de paridad
    - Para recuperar datos en un disco dañado, calcular XOR de bits desde otros discos (incluyendo el disco de bits de paridad)



Niveles de RAID (cont.)

- RAID de nivel 3 (Cont.)
  - Transferencia de datos más rápida que con un solo disco, pero menor E/S por segundo, dado que cada disco ha de participar en cada E/S.
  - Incluye el nivel 2 (aporta todas sus ventajas, a un coste menor).
- **RAID de nivel 4:** Paridad con bloques entrelazados; emplea distribución en el nivel de bloque y mantiene un bloque de paridad en un disco independiente para los correspondientes bloques de los otros *N* discos.
  - Cuando se graban bloques de datos, los bloques bits de paridad correspondientes se deben calcular y escribir sobre un disco de paridad
  - Para encontrar el valor de un bloque dañado, calcular XOR de bits desde los bloques correspondientes (incluyendo el bloque de paridad) de los otros discos.

(e) RAID 4: Paridad con bloques entrelazados

Niveles de RAID (cont.)

- RAID de nivel 4 (Cont.)
  - Aporta velocidades más altas de E/S, para lecturas de bloques independientes, que el nivel 3
    - la lectura de bloque se hace sobre un solo disco, de modo que los bloques almacenados en discos independientes se puedan leer en paralelo
  - Aporta altas velocidades de transferencia para lecturas de múltiples bloques no distribuidos
  - Antes de escribir un bloque se deben calcular los datos de paridad
    - Se puede hacer empleando bloques de paridad antiguos, valores viejos y nuevos del boque actual (2 bloques leídos + 2 bloques grabados)
    - O recalculando el valor de paridad, por medio de los valores nuevos de los bloques correspondientes al bloque de paridad
      - Más eficientes para las escrituras de grandes cantidades de datos secuenciales
  - El bloque de paridad se convierte en un cuello de botella para las escrituras de bloques independientes, dado que cada escritura de bloque también escribe sobre el disco de paridad

Niveles de RAID (cont.)

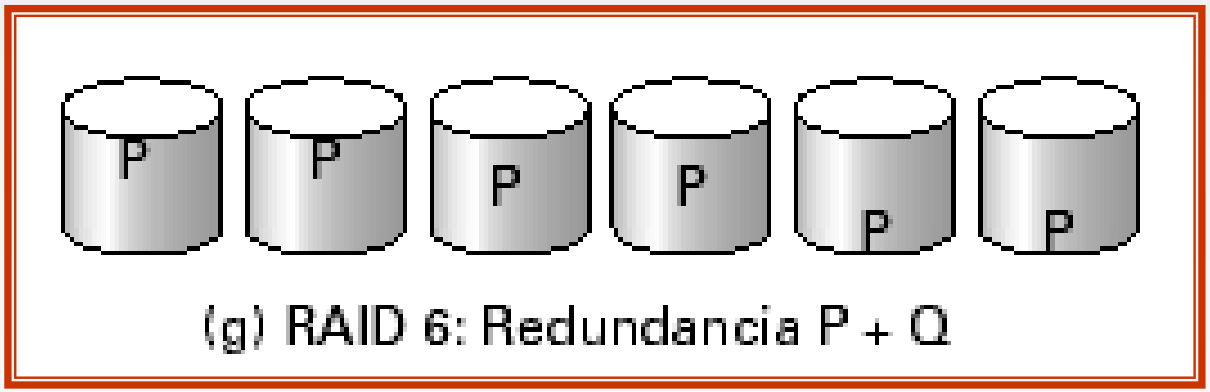
- **RAID de nivel 5:** Paridad distribuida con bloques entrelazados; datos y paridad divididos entre *N* + 1 discos, en vez de almacenar los datos en *N* discos y la paridad en 1..
  - Por ejemplo, con 5 discos el bloque de paridad para el *n*-ésimo conjunto de bloques se almacena en el disco (*n mod* 5) + 1, con los bloques de datos almacenados sobre los otros 4 discos.

(f) RAID 5: Paridad distribuida con bloques entrelazados

P0	0	1	2	3
4	P1	5	6	7
8	9	P2	10	11
12	13	14	P3	15
16	17	18	19	P4

Niveles de RAID (cont.)

- **RAID de nivel 5 (Cont.)**
  - Velocidades de E/S más altas que en el nivel 4.
    - Las escrituras de bloques tienen lugar en paralelo si los bloques y sus bloques de paridad están en discos diferentes.
  - Incluye el nivel 4: aporta algunas ventajas, pero evita los cuellos de botella del disco de paridad.
- **RAID de nivel 6**: Esquema de redundancia P+Q; similar al nivel 5, pero almacena información redundante para proteger contra fallos de los múltiples discos.
  - Mayor fiabilidad que en nivel 5 a un coste superior; no se usa ampliamente.



(g) RAID 6: Redundancia P + Q

Elección de los niveles de RAID

- Factores a tener en cuenta al seleccionar el nivel RAID
  - Coste económico
  - Rendimiento: Número de operaciones de E/S por segundo y ancho de banda durante la operativa normal
  - Rendimiento durante los fallos
  - Rendimiento durante la reconstrucción del disco fallido
    - Incluyendo el tiempo llevado en reconstruir el disco fallido
- RAID 0 sólo se usa cuando la seguridad de los datos no es importante
  - Por ejemplo, los datos se pueden recuperar rápidamente desde otras fuentes
- Los niveles 2 y 4 no se usan nunca dado que están incluidos en los niveles 3 y 5
- El nivel 3 no se usa más, dado que la distribución del bit hace que la lectura de un solo bloque obligue a acceder a todos los discos, gastando en el movimiento del brazo, cosa que evita la distribución de bloques (nivel 5)
- El nivel 6 apenas se emplea dado que los niveles 1 y 5 ofrecen una seguridad adecuada para la mayoría de las aplicaciones
- Así, la competencia solo está entre los niveles 1 y 5

Elección de los niveles de RAID (cont.)

- El nivel 1 aporta un rendimiento mucho mejor en escritura que el nivel 5
  - El nivel 5 requiere al menos 2 lecturas de bloques y 2 escrituras de bloques para grabar un solo bloque, mientras que el nivel 1 sólo requiere 2 escrituras de bloques
  - El nivel 1 es preferido en entornos de muchas actualizaciones, como en el disco del registro histórico
- El nivel 1 tenía un coste de almacenamiento superior que el nivel 5
  - las capacidades de las unidades de disco aumentan rápidamente (50% al año), mientras que el tiempo de acceso ha disminuido mucho menos (un factor 3 en 10 años)
  - Los requerimientos de E/S han aumentado mucho, por ejemplo, en los servidores Web
  - Cuando se han comprado discos suficientes para satisfacer la velocidad requerida de E/S, a menudo sobra capacidad de almacenamiento
    - ¡Por ello frecuentemente no hay coste monetario extra para el nivel 1!
- El nivel 5 es preferido para aplicaciones con velocidad de actualización baja grandes cantidades de datos
- El nivel 1 se prefiere para todas las otras aplicaciones

Aspectos del hardware

- **Software RAID**: Las implantaciones RAID se hacen totalmente en software, sin ningún soporte hardware especial
- **Hardware RAID**: Implantaciones RAID con hardware especial
  - Se emplea RAM no volátil para registrar las escrituras que se están ejecutando
  - Tener cuidado con: fallos en el suministro eléctrico durante la escritura pueden originar la corrupción del disco
    - Por ejemplo, fallos después de escribir un bloque, pero antes de escribir el segundo en un sistema de imagen
    - Así, los datos corruptos deben detectarse cuando se reanuda el suministro eléctrico
      - La recuperación de la corrupción es similar a la recuperación desde discos fallidos
      - NV-RAM ayuda a detectar de manera eficiente bloques potencialmente corruptos
        - De lo contrario, todos los bloques del disco deben leerse y compararse con los bloques espejo/paridad

Aspectos del hardware (cont.)

- **Intercambio en caliente**: sustitución del disco mientras está funcionando, sin cortar el suministro eléctrico
  - Soportado por algunos sistemas de hardware RAID,
  - reduce el tiempo de recuperación y mejora enormemente la fiabilidad
- Muchos sistemas mantienen discos de recambio, que se mantienen en línea y se usan para reemplazar los discos fallidos inmediatamente que se detecta el fallo
  - Muchos sistemas mantienen discos de recambio, que se mantienen en línea y se usan para reemplazar los discos fallidos inmediatamente que se detecta el fallo
- Muchos sistemas de hardware RAID aseguran que un solo punto de fallo no detendrá el funcionamiento del sistema, empleando
  - Fuentes de alimentación redundantes con batería de salvaguarda
  - Múltiples controladores e interconexiones, como protección contra fallos de controlador/interconexión

Discos ópticos

- Discos compactos con memoria de sólo lectura (CD-ROM)
  - Los discos pueden cargarse en, o eliminarse de una unidad
  - Elevada capacidad de almacenamiento (640 MB por disco)
  - Elevado tiempo de búsqueda de aproximadamente 100 milisegundos (la cabeza de lectura óptica es más pesada y lenta)
  - Latencia más alta (3.000 RPM) y menor velocidad de transferencia de datos (3-6 MB/s), comparada con los discos magnéticos
- Video disco digital (DVD)
  - El DVD-5 almacena 4.7 GB y el DVD-9 8.5 GB
  - DVD-10 y DVD-18 están formateados por las dos caras, con capacidades de 9.4 GB y 17 GB
  - Otras características similares al CD-ROM
- Las versiones de grabación de una sola vez (CD-R y DVD-R) se están haciendo populares
  - los datos sólo se pueden escribir una vez y no se pueden borrar.
  - alta capacidad y larga vida; se usan para el almacenamiento de archivos
  - Versiones de escritura múltiple (CD-RW, DVD-RW, DVD+RW y DVD-RAM) también están disponibles

Cintas magnéticas

- Contienen grandes volúmenes de datos y aportan velocidades de transferencia altas
  - Unos pocos GB para el formato DAT (Cinta de audio digital), 10-40 GB con formato DLT (Cinta lineal digital), más de 100 GB con formato Ultrium y 330 GB con formato de exploración helicoidal Ampex
  - Velocidades de transferencia desde unos pocos hasta 10 MB/s
- Actualmente es el medio de almacenamiento más barato
  - Las cintas son baratas, pero el coste de las unidades es muy alto
- Tiempo de acceso muy lento, en comparación con los discos magnéticos y ópticos
  - limitado a accesos secuenciales.
  - Algunos formatos (Accelis) soportan búsquedas más rápidas (décimas de segundo) al precio de reducir la capacidad
- Usado principalmente para copias de seguridad, para el almacenamiento de información que se usa poco frecuentemente y como un medio sin conexión para la transferencia de información desde un sistema a otro.
- Los cambiadores de cintas se emplean para el almacenamiento de muy alta capacidad
  - desde terabyte (10^{12} bytes) hasta petabyte (10^{15} bytes)

Acceso al almacenamiento

- Un archivo de base de datos está dividido en unidades de almacenamiento de longitud fija, denominadas **bloques**. Los bloques son unidades de asignación de almacenamiento y de transferencia de datos.
- El sistema de bases de datos busca minimizar el número de transferencias de bloques entre el disco y la memoria. Se puede reducir el número de accesos a disco manteniendo en memoria tantos bloques como sea posible.
- **Memoria intermedia** – parte de la memoria principal disponible para almacenar copias de bloques del disco.
- **Gestor de la memoria intermedia** – subsistema responsable de asignar el espacio de la memoria intermedia en la memoria principal.

Gestor de la memoria intermedia

- Los programas llaman al gestor de memoria intermedia cuando necesitan un bloque del disco.
  1. Si el bloque ya está en la memoria intermedia, a la solicitud del programa se da la dirección del bloque en la memoria principal
  - Si el bloque no está en la memoria intermedia, el gestor de memoria intermedia
    1. asigna espacio en la memoria intermedia para el bloque
       1. reemplazando (desechando) algún otro bloque, si es necesario, para hacer espacio al nuevo bloque.
       2. El bloque desechado se graba de nuevo a disco, sólo si se modificó desde el momento en que fue grabado a / tomado del disco.
    2. lee el bloque desde el disco a la memoria intermedia y pasa la dirección del bloque en la memoria principal al solicitante.

Políticas de sustitución de la memoria intermedia

- La mayoría de los sistemas operativos reemplazan el bloque **menos recientemente utilizado** (estrategia LRU)
- Idea tras LRU – utilizar el último modelo de referencias del bloque como un indicador de referencias futuras
- Las consultas han de definir bien los modelos de acceso (tales como búsquedas secuenciales) y un sistema de base de datos puede utilizar la información de una consulta de usuario para predecir referencias futuras
  - LRU puede ser una mala estrategia para ciertos modelos de accesos que implican búsquedas repetidas de datos
    - por ejemplo, al calcular la reunión de 2 relaciones *r* y *s* mediante un bucle anidado
      para cada tupla *tr* de *r* hacer
        para cada tupla *ts* de *s* hacer
          si las tuplas *tr* y *ts* se corresponden …
  - Es preferible una estrategia mixta, con sugerencias sobre la estrategia de sustitución aportada por el optimizador de consultas

Políticas de sustitución de la memoria intermedia (cont.)

- **Bloque clavado** – bloque de memoria que no tiene permitido ser grabado de nuevo a disco.
- **Estrategia de extracción inmediata** – libera el espacio ocupado por un bloque tan pronto como se procesa la tupla final de ese bloque
- Estrategia del utilizado más recientemente (MRU) – el sistema debe clavar el bloque que se está procesando actualmente. Después de procesar la tupla final de ese bloque, se desclava y se convierte en el bloque más recientemente utilizado.
- El gestor de la memoria intermedia puede utilizar información estadística con respecto a la probabilidad de que una petición referencie una determinada relación
  - Por ejemplo, el diccionario de datos es accedido frecuentemente. Heurística: mantiene los bloques del diccionario de datos en la memoria intermedia de la memoria principal
- Los gestores de la memoria intermedia también soportan la salida forzada de bloques con fines de recuperación (más en el Capítulo 17)

Organización de archivos

- La base de datos está almacenada como un grupo de *archivos*. Cada archivo es una secuencia de *registros*. Un registro es una secuencia de *campos*.
- Un enfoque puede ser:
  - suponer que el tamaño del registro es fijo
  - cada archivo tiene sólo registros de un determinado tipo
  - se usan diferentes archivos para diferentes relaciones

  Este caso es el más fácil de implementar; se considerarán registros de longitud variable posteriormente.

Registros de longitud fija

- Enfoque sencillo:
  - Almacenar el registro *i* empezando desde el byte $n * (i - 1)$, donde *n* es el tamaño de cada registro.
  - El acceso al registro es sencillo, pero los registros pueden atravesar bloques
    - Modificación: no se permite a los registros atravesar los límites de un bloque
- Borrado del registro *I:* alternativas*:*
  - mover los registros $i + 1, \ldots, n$ a $i, \ldots, n - 1$
  - mover el registro *n* a *i*
  - no mover registros, sino enlazar todos los registros libres sobre una *lista libre*

registro 0	C-102	Navacerrada	400
registro 1	C-305	Collado Mediano	350
registro 2	C-215	Becerril	700
registro 3	C-101	Centro	500
registro 4	C-222	Moralzarzal	700
registro 5	C-201	Navacerrada	900
registro 6	C-217	Galapagar	750
registro 7	C-110	Centro	600
registro 8	C-218	Navacerrada	700

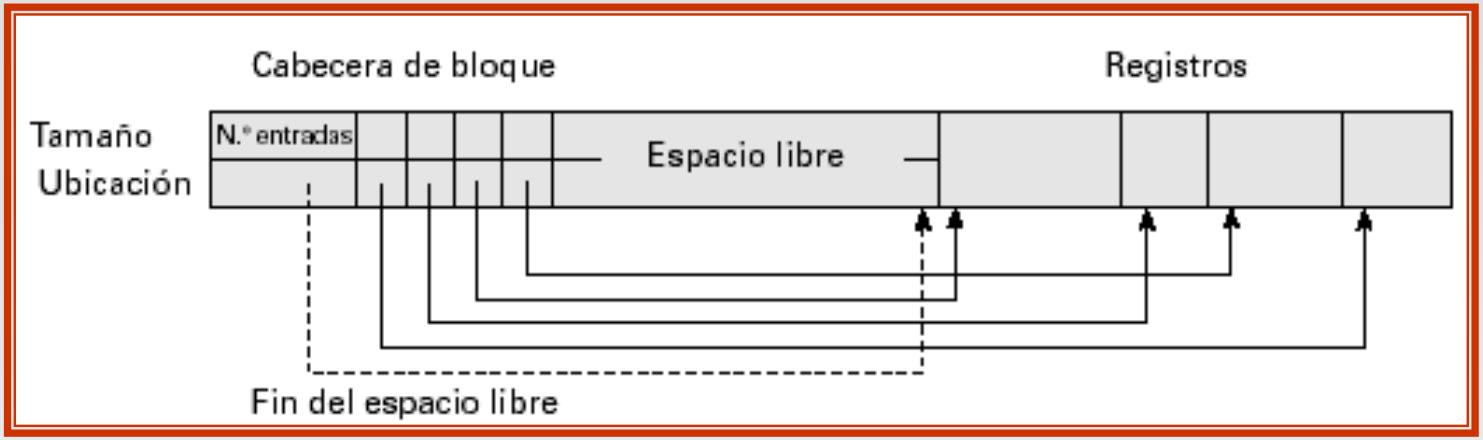
Listas libres

- Almacenar la dirección del primer registro borrado en la cabecera del archivo.
- Emplear este primer registro para almacenar la dirección del segundo registro borrado, etcétera
- Se puede pensar en estas direcciones como punteros dado que “apuntan” a la posición de un registro.
- Representación más eficiente del espacio: reutilización del espacio para atributos normales de registros libres para almacenar punteros. (No se almacena ningún puntero en los registros en uso.)

Registros de longitud variable

- Los registros de longitud variable surgen en los sistemas de bases de datos de diferentes maneras:
  - Almacenamiento en un archivo de múltiples tipos de registro.
  - Tipos de registro que permiten longitudes variables de uno más campos.
  - Tipos de registro que permiten campos repetidos (empleados en algunos de los más antiguos modelos de datos).

Registros de longitud variable: Estructura de páginas con ranuras



- La cabecera de las páginas con ranuras contiene:
  - número de entradas del registro
  - final del espacio libre en el bloque
  - localización y tamaño de cada registro
- Los registros se pueden mover alrededor de una página para mantenerlos contiguos, sin espacio vacío entre ellos; se debe actualizar la entrada en la cabecera.
- Los punteros no deberían apuntar directamente al registro — en su lugar, deberían apuntar a la entrada para el registro en la cabecera.

Organización de registros en archivos

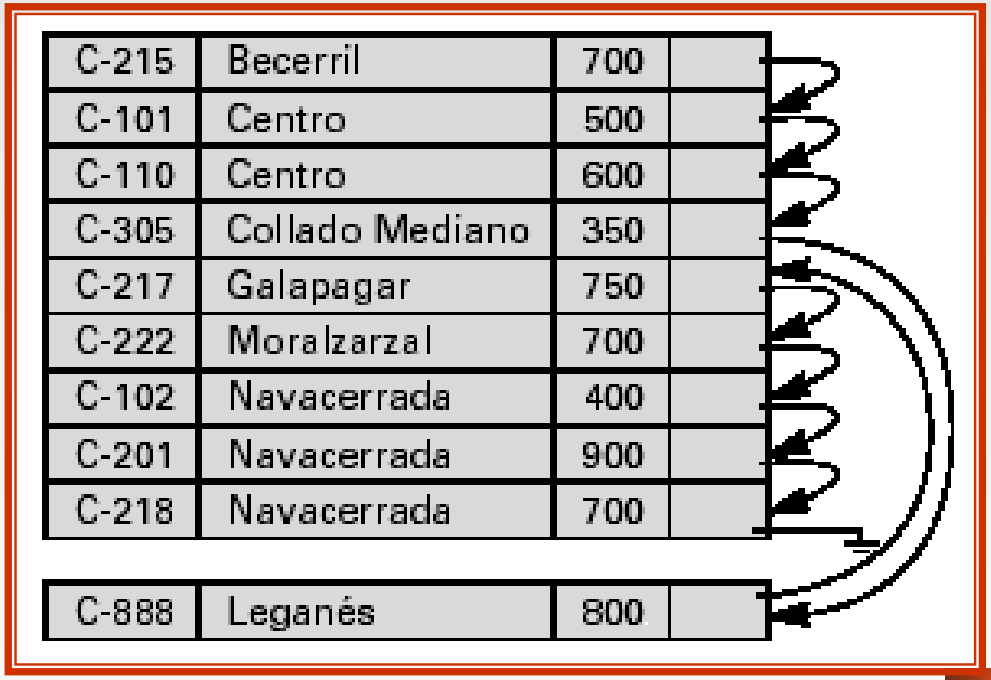
- **Pila** – un registro puede estar colocado en cualquier parte del archivo donde haya espacio
- **Secuencial** – almacena registros en orden secuencial, de acuerdo al valor de la clave de búsqueda de cada registro
- **Asociación** – una función de asociación calculada sobre algún atributo de cada registro; el resultado determina en qué bloque del archivo se debería situar el registro
- Los registros de cada relación pueden almacenarse en un archivo independiente. En una **organización de archivos en agrupaciones** los registros de diferentes relaciones se pueden almacenar en el mismo archivo
  - Motivación: almacenar registros relacionados en el mismo bloque minimiza la E/S

Organización de archivos secuencial

- Adecuada para aplicaciones que requieren el procesamiento secuencial de todo el archivo
- Los registros del archivo están ordenados por una clave de búsqueda

Organización de archivos secuencial (Cont.)

- Borrado – emplear cadenas de punteros
- Inserción –localizar la posición donde se va a insertar el registro
  - si hay espacio libre insertarlo ahí
  - si no hay espacio libre, insertar el registro en un bloque de desbordamiento
  - En cualquier caso, la cadena de punteros debe actualizarse
- Necesidad de reorganizar el archivo periódicamente para restablecer el orden secuencial



Organización de archivos en agrupaciones de varias tablas

Almacenar varias relaciones en un archivo utilizando una organización del archivo en **agrupaciones de varias tablas**

nombre_cliente	*número_cuenta*
López	C-102
López	C-220
López	C-503
Abril	C-305

nombre_cliente	*calle_cliente*	*ciudad_cliente*
López	Mayor	Arganzuela
Abril	Preciados	Valsaín

Organización de archivos en agrupaciones de varias tablas (cont.)

Organización de archivos en agrupaciones de varias tablas de *cliente* e *impositor*

López	Mayor	Arganzuela
López	C-102	
López	C-220	
López	C-503	
Abril	Preciados	Valsaín
Abril	C-305	

- Bueno para consultas que impliquen *impositor* ⋈ *cliente* , y para consultas que impliquen un único cliente y sus cuentas.
- Malo para consultas que impliquen a un único cliente
- Los resultados en registros de longitud variable
- Pueden añadir cadenas de punteros para enlazar registros de una determinada relación.

Almacenamiento del diccionario de datos

El diccionario de datos (también denominado catálogo del sistema) almacena metadatos: es decir, datos acerca de datos, como

- Información sobre relaciones
  - nombres de relaciones
  - nombres y tipos de atributos de cada relación
  - nombres y definiciones de vistas
  - restricciones de integridad
- Información del usuario y de la cuenta, incluyendo contraseñas
- Datos estadísticos y descriptivos
  - número de tuplas en cada relación
- Información de la organización del archivo físico
  - Como está almacenada la relación (secuencial/asociativa/ ....)
  - Localización física de la relación
- Información sobre índices (Capítulo 12)

Almacenamiento del diccionario de datos (cont.)

- Estructura del catálogo:
  - Representación relacional en el disco
  - Estructuras de datos especializadas diseñadas para acceso eficiente, en memoria
- Una posible representación del catálogo:

Relación_metadato = (nombre_relación, número_de_atributos, organización_almacenamiento, ubicación)
Atributo_metadato = (nombre_atributo, nombre_relación, tipo_dominio, posición, longitud)
Usuario_metadato = (nombre_usuario, contraseña_cifrada, grupo)
Índice_metadato = (nombre_índice, nombre_relación, tipo_índice, atributos_índice)
Vista_metadato = (nombre_vista, definición)

Fin del capítulo 11

Fundamentos de Bases de datos, 5ª Edición.