Organización y manejo de archivos

www.inf.ucv.cl/ wpalma/oma

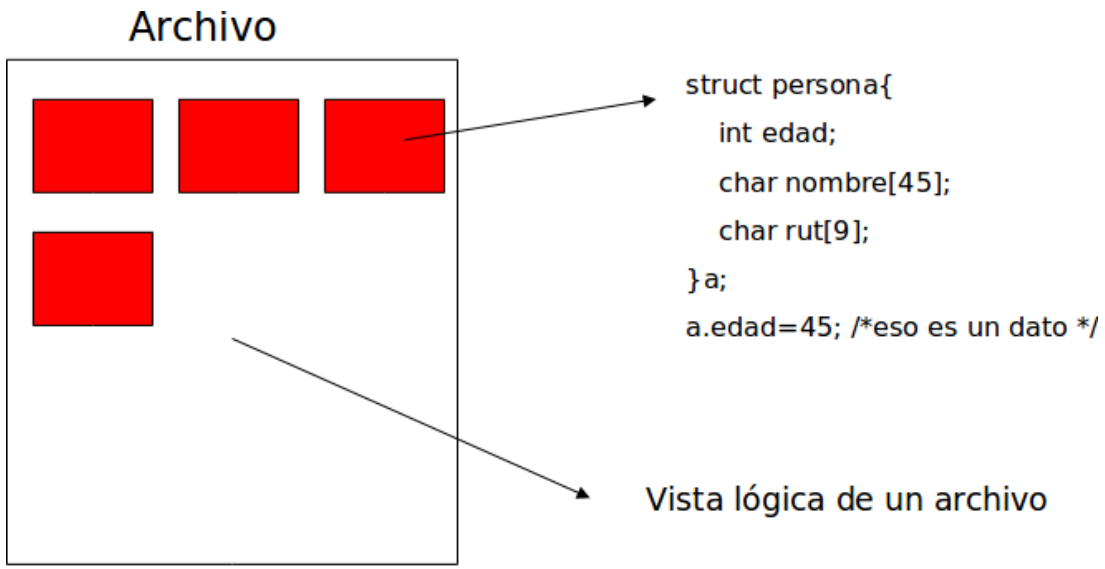
Dr. Wenceslao Palma
wenceslao.palma@ucv.cl

Conceptos

- Los datos son lo más importante para una organización.
- Se constituyen en el elemento fundamental para el proceso de toma de decisiones.
- Un proceso de administración de datos debe contemplar
  - Acceso selectivo y eficiente.
  - Procesamiento y presentación.
  - Garantía en consistencia de los datos.

Conceptos

- Archivo: generalmente corresponde a una colección de registros lógicamente relacionados.
- Registro: es una estructura de campos o de datos lógicamente relacionados. Generalmente un registro se puede identificar de manera única.
- Dato: representación de un atributo en el dominio de un problema.



Conceptos

La técnica utilizada para representar y almacenar registros es llamada organización de archivos. Todo tipo de organización de archivos busca la conveniencia respecto de tiempos de acceso y recuperación. La organización más apropiada para un archivo dependerá de las características operacionales del medio de almacenamiento y de la naturaleza de las operaciones sobre los registros.

Conceptos

Considerando el almacenamiento se tiene la siguiente jerarquía: Memoria Caché, Memoria Principal, Discos Magnéticos, Memoria terciaria. En forma ideal siempre es recomendable trabajar con los datos en memoria volátil, debido a la velocidad de acceso, para luego enviarlos a memoria secundaria o terciaria.

- Memoria Caché
  - su objetivo es reducir los tiempos de espera.
  - es pequeña pero de mucha rápidez.
  - es necesario administrarla.
- Memoria Principal
  - se utiliza para el procesamiento de los datos.
- Disco Magnético
  - no es volátil.
  - dispositivo de acceso aleatorio.
  - permite almacenar gran cantidad de bytes.
  - para procesar su contenido es necesario el uso de la memoria principal.

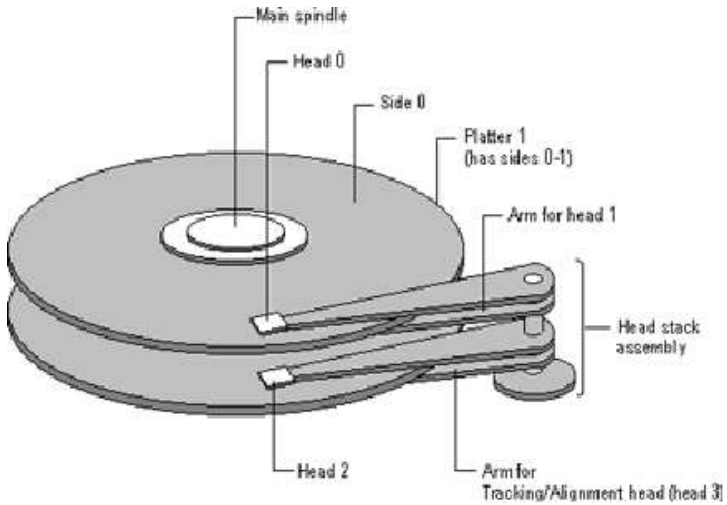
Conceptos

- Memoria terciaria
  - considera entre otros, CDs, DVDs y Cintas.
  - una cinta es un dispositivo de acceso secuencial, es mas lenta que un disco magnético, generalmente se utiliza para respaldos.

La principal desventaja en la utilización de Discos se encuentra en el tiempo de acceso y de recuperación. Sin embargo este se puede disminuir seleccionando una organización adecuada.

Disco magnético

también conocido como disco duro, se compone de un pack de platos cada uno de los cuales tiene caras, pistas y sectores para almacenar la información. Además posee cabezas, asociadas a un brazo, para leer y grabarla.



Una cabeza se mueve radialmente sobre la superficie de un plato que gira a gran velocidad. Sólo una cabeza puede realizar transferencia de datos en un determinado momento.

Conceptos

El tiempo de acceso a un disco está dado por tiempo de seek+latencia (rotacional)+ tiempo de transferencia.

- seek: tiempo requerido para transitar entre pistas.
- latencia: tiempo requerido para que el disco gire al sector que se desea. Generalmente corresponde a media vuelta.
- transferencia: velocidad con la que se escriben/leen bytes hacia/desde el disco.

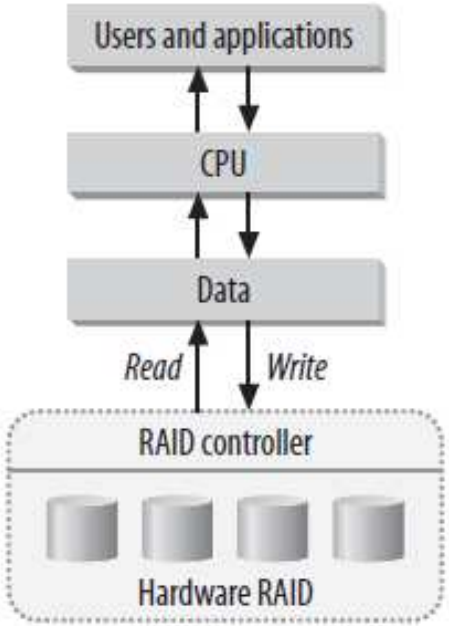
Ejercicio: considere el ordenamiento y mezcla de un archivo de 800000 registros. Sólo 10000 se pueden almacenar en RAM. Si el tamaño del registro es 100 bytes y se utiliza un disco de 3600 RPM que posee seek de 18 ms y transferencia de 1229 bytes/ms.
Cuanto es el tiempo requerido para realizar el ordenamiento y mezcla?

Conceptos

- Para mejorar el desempeño
  - minimizar movimientos del brazo.
  - lo ideal para la latencia es que la cabeza se encuentre justo en el sector que contiene el dato que se requiere. Sectores contiguos v/s sectores intercalados.
  - transferencia en paralelo desde el pack de discos, para esto se requieren varios brazos. Ante esto nace el concepto de RAID (Redundant Array of Independent Disks).

RAID (Redundant Array of Independent Disks)

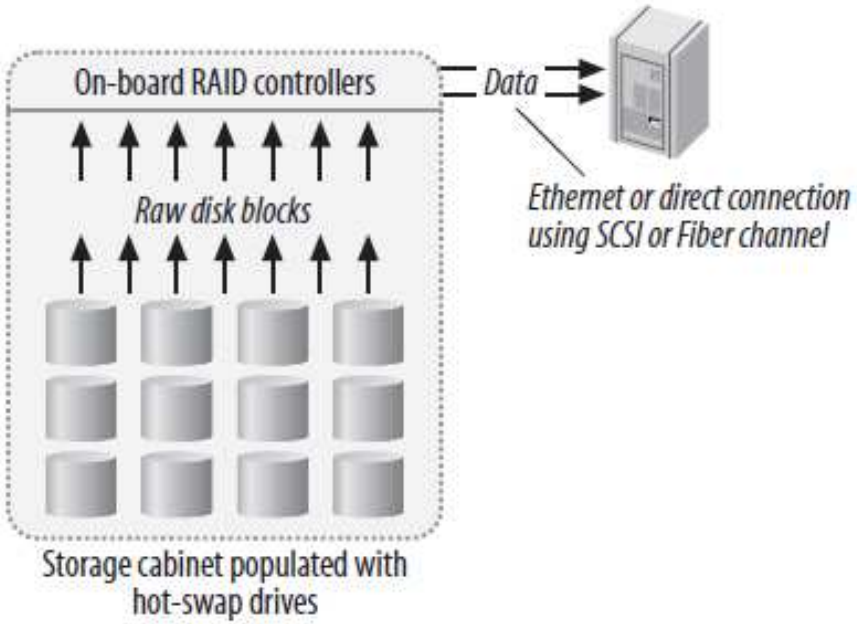
- El array es administrado por una controladora que contiene RAID firmware.
- RAID administrado por hardware es más rápido que el administrado por software.



- La controladora posee una BIOS que proporciona las herramientas de administración para la configuración y mantención.
- El sistema operativo ve el arreglo como un “gran disco duro”.

RAID (Redundant Array of Independent Disks)

También existen soluciones externas, las cuales son conectadas mediante SCSI, Ethernet o Fibra.



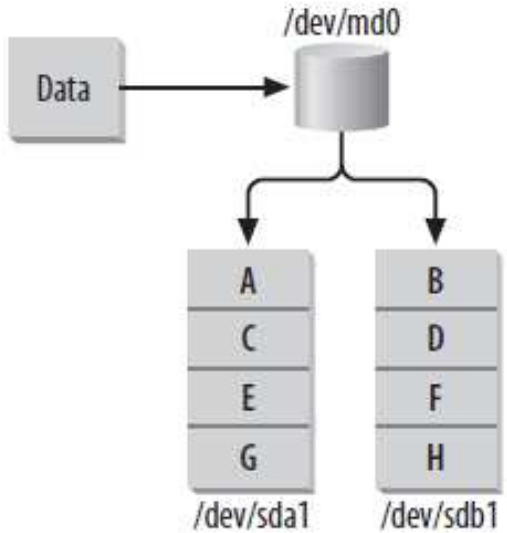
RAID (Redundant Array of Independent Disks)

Niveles RAID

Existen distintas alternativas de configurar un arreglo de discos. Generalmente un nivel se define en base a los requerimientos de aplicaciones que lo utilizan. Básicamente constituyen un compromiso entre redundancia y rendimiento.

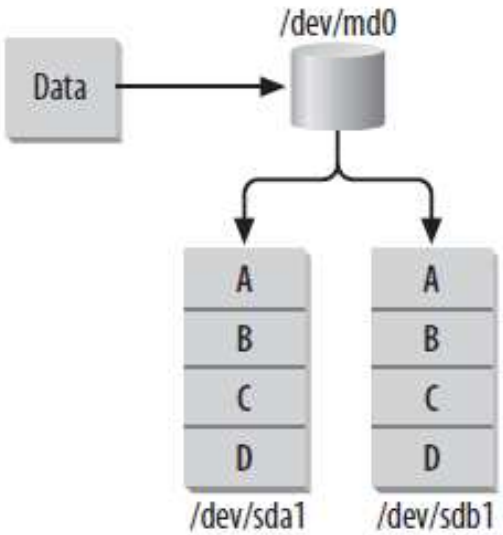
RAID 0(striping)

En rigor no corresponde a un nivel RAID, debido a que no presenta redundancia. Conviene utilizarlo cuando los datos no son críticos y el rendimiento es importante.



RAID 1(mirroring)

Proporciona la forma más completa en cuanto redundancia. Puede soportar muchas fallas sin la necesidad de un algoritmo de recuperación.

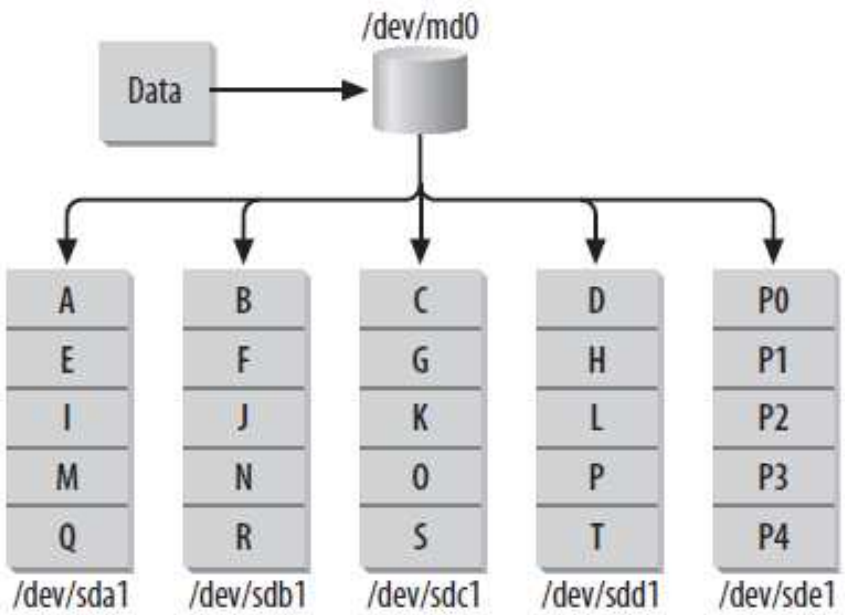


RAID 1(mirroring)

El rendimiento de las operaciones de escritura es inversamente proporcional al aumento de discos en el array. Por otro es posible realizar lecturas en forma concurrente.
IMPORTANTE: el costo de implementación es como mínimo el doble del requerimiento de almacenamiento.

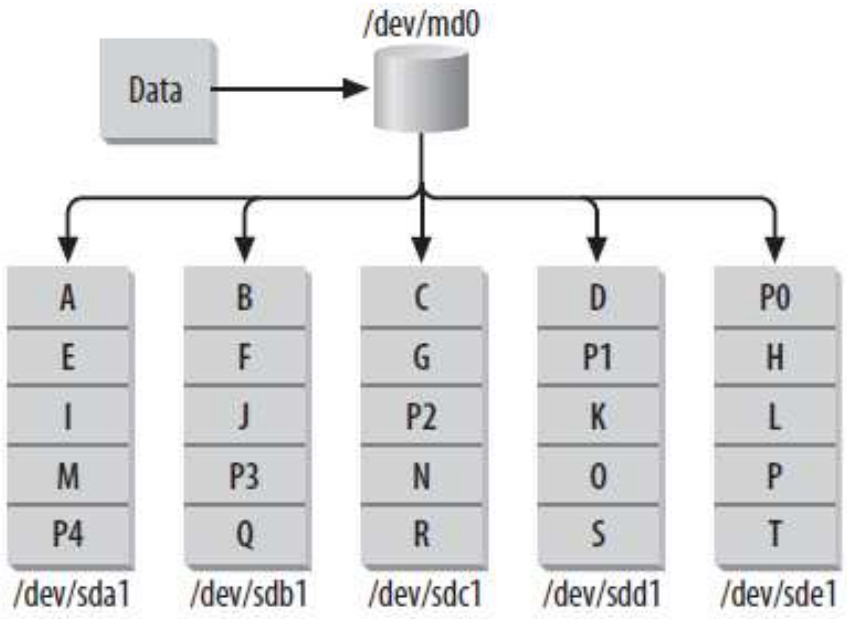
RAID 4

Utiliza un disco para almacenar información de paridad, la cual se utiliza ante un eventual desastre.

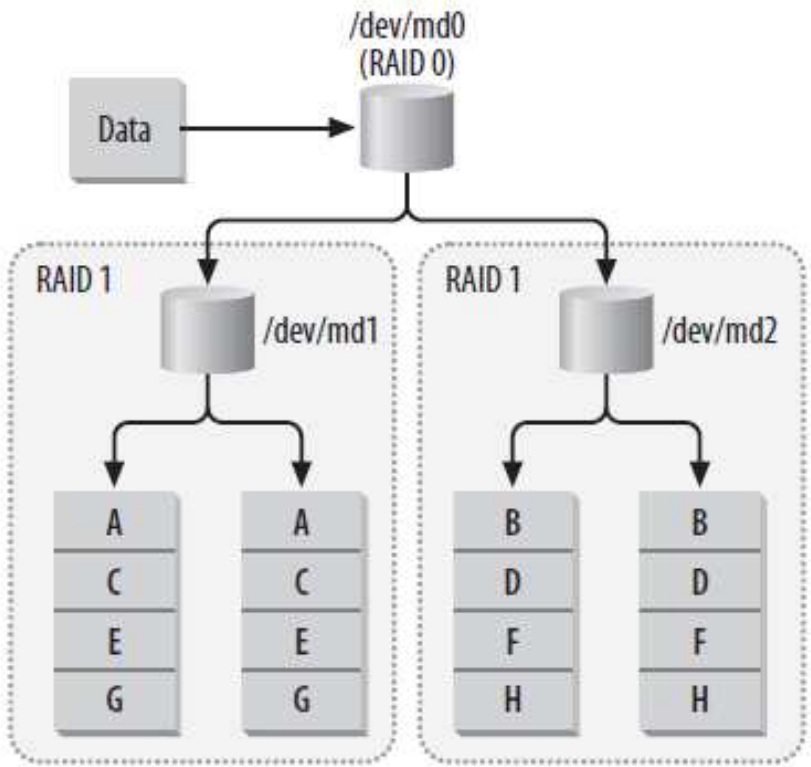


RAID 5

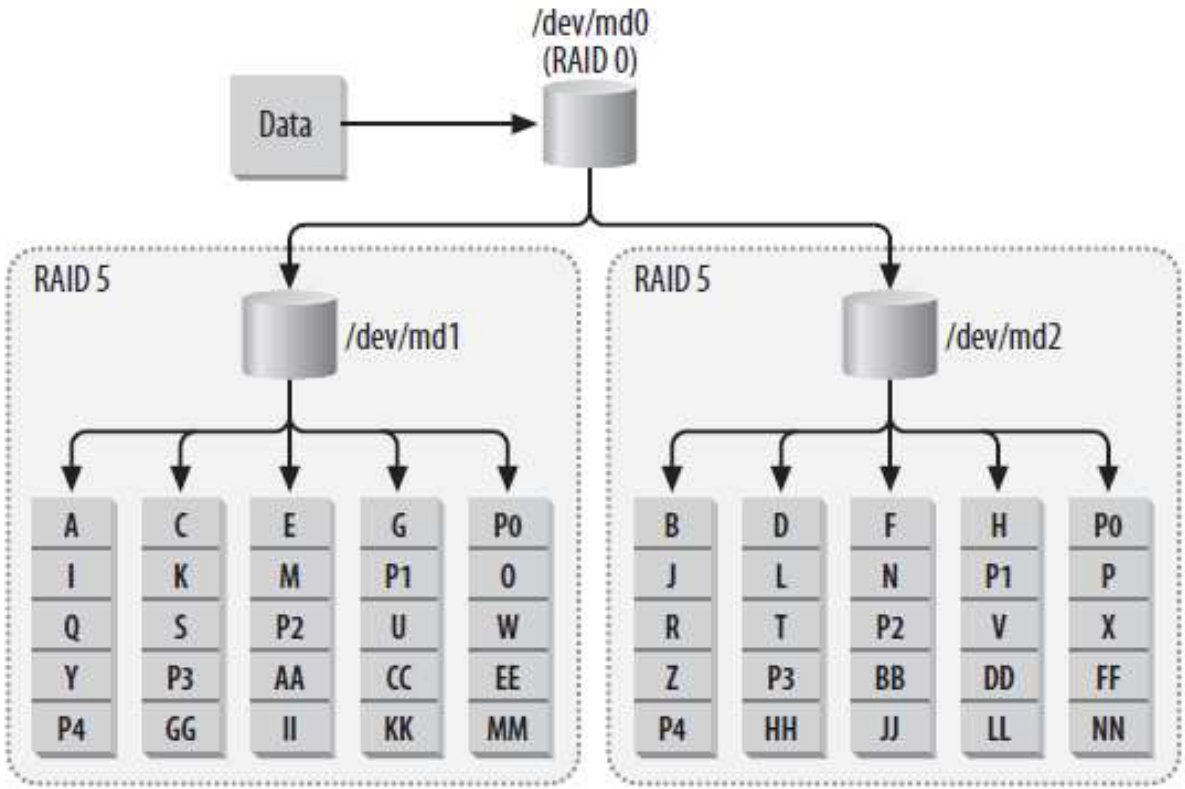
Aquí se elimina el uso de un disco exclusivo para el tema de la paridad. Se utiliza un bloque por cada disco



RAID 10 (striping mirror)



RAID 50 (striping parity)



RAID (Redundant Array of Independent Disks)

	RAID 1	RAID 0	RAID 1	RAID 5
Escritura	Lenta, mas aún si se agregan discos	mejor que un único disco	comparable a RAID con un disco menos	comparable a RAID con un disco menos
Lectura	Rápida, mas aún si se agregan discos	el mejor	comparable a RAID con un disco menos	comparable a RAID con un disco menos
Nro fallas	n-1	0	1	1
Aplicaciones	Servidor de video, imágenes. En gnral sistemas con poca volatilidad	uso doméstico	Servidor de archivos, BD	Equivalente a RAID 5

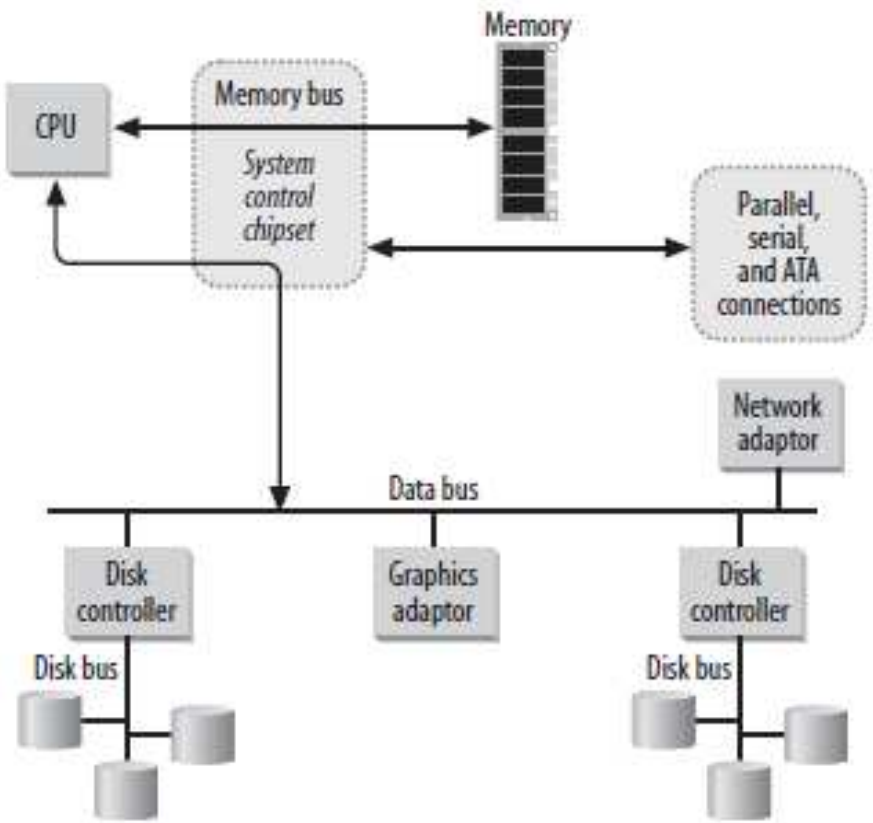
Consideraciones de hardware

No debemos olvidar que un array es solo un componente mas de un sistema computacional. Muchos factores afectan el rendimiento y la capacidad de expansión de un arreglo de discos:

- Throughput del bus.
- Canales de I/O.
- Throughput del protocolo de acceso al disco.
- Velocidad del disco.
- CPU y memoria.

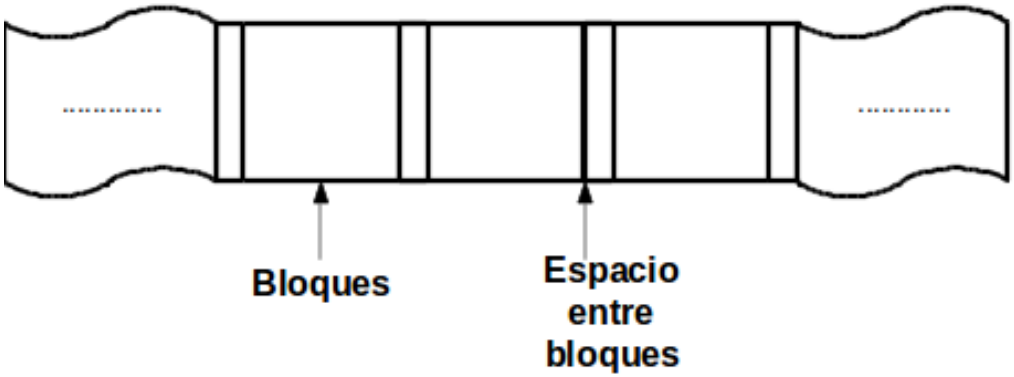
Considerando estos factores la idea es poner atención a los posibles cuellos de botella que se puedan generar.

RAID (Redundant Array of Independent Disks)



La velocidad de los buses de datos y de disco tiene un impacto directo en el desempeño del sistema. Es sencillo agregar más controladoras de disco aumentando el throughput pero el bus de datos es uno solo.

Cintas



Dispositivo de almacenamiento secuencial.
Parámetros relevantes

- Espacio entre bloques
- Largo de la cinta
- Densidad
- tamaño de los registros
- factor de bloqueo

Ejemplo

Largo : 2400 [pies] (1 pie = 12 pulgadas), Densidad: 6250 [bpp], Tamaño registro: 100 bits, Factor de bloqueo (fb): 1, Espacio entre bloques: 0.6 [pulg]
Cuántos bloques puede contener la cinta?
Cuál es la eficiencia en el uso de la cinta?

Tamaño de un bloque (recordar fb=1)

1[pulg] $\rightarrow$ 6255 [bits]
X $\rightarrow$ 100 [bits]
X = 0.016 [pulg]

Cantidad de bloques

Largo de la cinta: 2400 [pies] * 12 [pulg/pie] = 28800 [pulg]
Cantidad de bloques: 28800 [pulg]/(0.016 + 0.6)[pulg] = 46753 [bloques]

Eficiencia en el uso de la cinta

Debemos calcular el espacio ocupado por los datos.
Cuando el fb es 1 : 46753*0.016 = 748 [pulg] $\rightarrow$ 2.59%

Supongamos fb=20

X = 0.32 [pulg]
La cantidad de bloques será: 28800 [pulg]/(0.32 + 0.6)[pulg] = 31304 [bloques]
Y la eficiencia : 31304*0.32 = 10017 [pulg] $\rightarrow$ 34.78%

Analicemos el tiempo de respuesta cuando se lee toda la cinta (tiempo para leer bloques + detención entre espacios)

Supongamos: Acceso lectura/escritura: 200 [pulg/seg], Detención entre espacios: 4 [ms]
Si fb es 1 : 3.74 [seg] + 187.01 [seg] = 190.75 [seg]
Si fb es 20: 50.08 [seg] + 125.22 [seg] = 175.3 [seg]

Respaldo y recuperación

- Ante pérdida o daño en los datos (medio de almacenamiento) es necesaria una política de respaldo y recuperación.
- Para una recuperación : Respaldo, Journaling, Checkpoint.

Respaldo

Corresponde a una copia de los datos. Periodicidad: diaria, semanal, mensual. (cold backup, hot backup)

Journaling

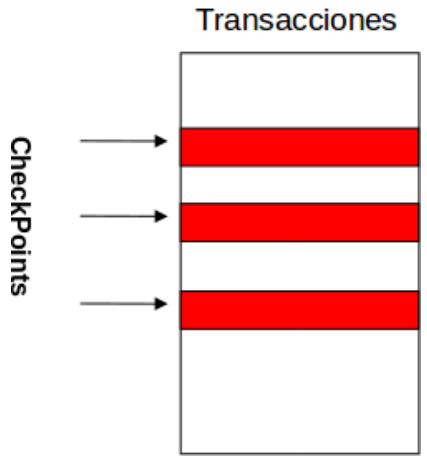
Corresponde a un registro (log) de los procesos (transacciones) y las actualizaciones realizadas sobre los datos.

Journaling

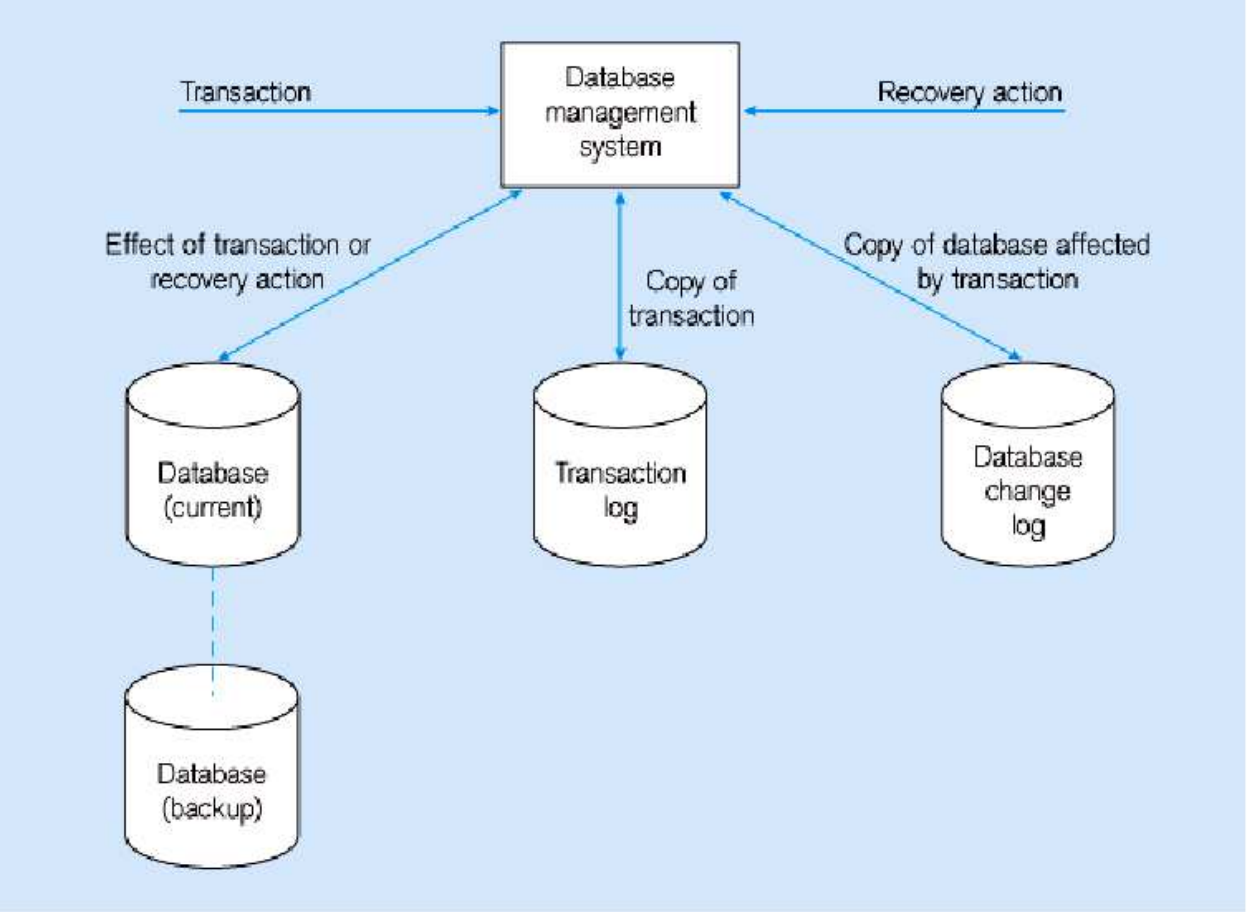
- Transactional Log: registro de acciones realizadas por las transacciones.
- Database Change Log : imagen de los datos actualizados.
  - Before-image: copia antes de la modificación.
  - After-image: copia después de la modificación.

Checkpoint

Cuando se actualiza un dato vale la pena guardar un registro especial en el log de transacciones y pre-imágenes. Util para realizar la recuperación desde la "última marca".

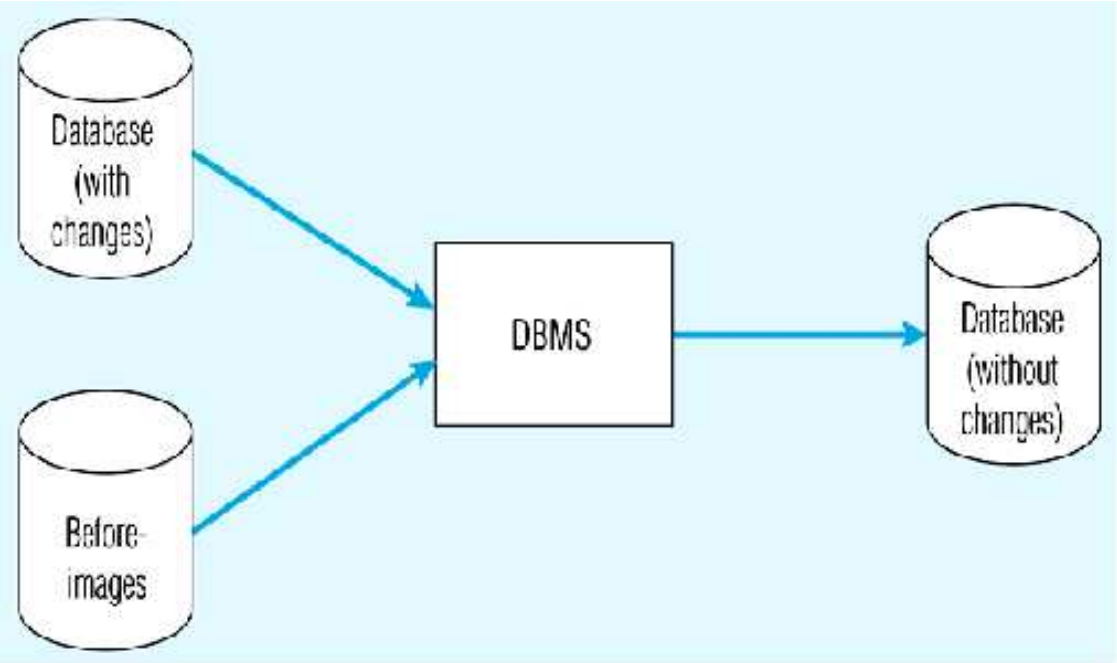


BD y recuperación



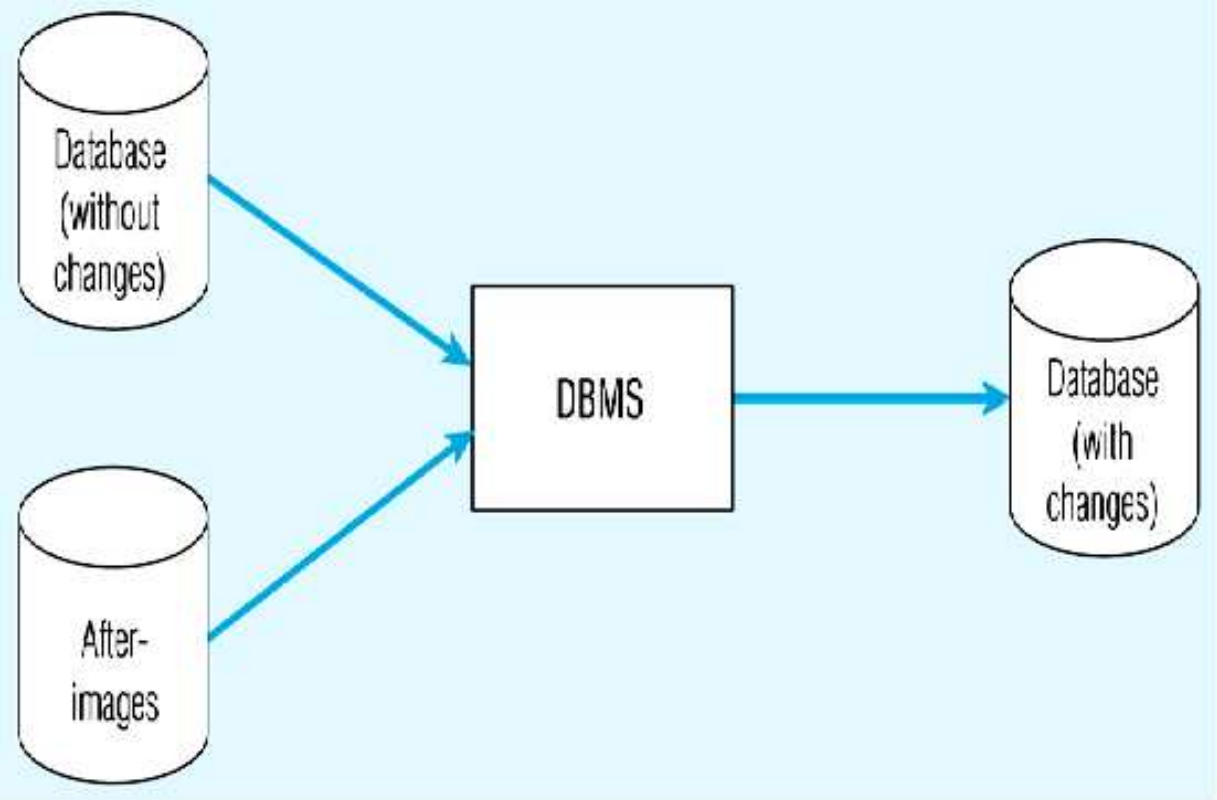
Procedimientos de recuperación

- restore-rerun: reprocesar transacciones contra el backup.
- rollback: se aplica la pre-imagen



Procedimientos de recuperación

- rollforward



Consistencia

Por el momento nos interesa la noción de consistencia. Consideremos el siguiente procedimiento

```
procedure deposit
begin
      start
      input(account#, amount);
      temp:= read(accounts[account#]);
      temp:= temp+amount;
      write(accounts[account#], temp);
      commit
end;
```

Y las siguientes acciones de dos clientes (cliente1 y cliente2):

- Ambos clientes realizan depósitos sobre la misma cuenta account13, la cual posee inicialmente $1000.
- cliente1 deposita $100
- "Al mismo tiempo" cliente2 deposita $100000

Consistencia

Se origina la siguiente ejecución concurrente:

```
read1(accounts[13]);
read2(accounts[13]);
write2(accounts[13],$101000);
commit2;
write1(accounts[13],$1100);
commit1;
```

El depósito del cliente2 se perdió!!!!
Es necesario garantizar consistencia de los datos ante ejecuciones concurrentes.
Idea: permitir concurrencia siempre y cuando una ejecución intercalada de 2 o más procesos tenga el mismo efecto que una ejecución serial. Lo anterior se conoce como **ejecución serializable**.