FICHEROS

Javier Fernández Rivera - www.aurea.es

Introducción

Ficheros o Archivos: Los ficheros son unas unidades lógicas de almacenamiento que define el propio sistema operativo y cuyo significado esta definido por su creador. Los ficheros están constituidos a nivel interno por un conjunto de registros lógicos.

Pongamos el ejemplo de un fichero grafico, de una escala de grises. Este archivo seria un conjunto de bits definidos por el creador, y que internamente seria una matriz numérica, cuyos elementos representan los niveles de grises de cada uno de los píxeles de la imagen que contiene el fichero.

Los ficheros o archivos se identifican por su nombre y su extensión. Según que sistema operativo se utilice, podemos introducir un mayor o menor numero de caracteres a este nombre. Por ejemplo, MSDOS, solo permite 8 caracteres para el nombre, de hay que si en Windows ficheros con un nombre de mas de 8 caracteres, MSDOS, los reemplaza por "~1" ((Alt. +126) 1). En Windows, en realidad como sistema operativo solo hace un apaño para esto, identifica solo los primeros caracteres, el resto de nombre de un fichero o archivo lo añade pero no lo identifica con él, es un apaño XD. En otros sistemas operativos como LINUX si permite la introducción de un nombre para ficheros con mas de 8 caracteres y si los identifica con cuantos sean.

Los ficheros almacenan dentro de ellos sus propias características, como son: la fecha de su creación, la fecha de su ultima modificación, sus atributos (solo lectura, etc.), su tamaño, etc.

Tipos y estructuras de los ficheros/archivos: La extensión que era la segunda parte que identificaba a un fichero, es la parte que define el tipo de fichero de que se trata. Así pues si tenemos un fichero con extensión: jpg, bmp, gif, estaremos hablando de un fichero de imagen. Si es con extensión doc, será de documentos, si es: exe (ejecutable), com (de comandos), bat (procesamiento por lotes (programación lineal)"baths") ficheros ejecutables.

Definiciones

Registro lógico (record): Es el conjunto de datos referentes a una misma entidad que constituye una unidad para un determinado proceso ejecutable (programa o parte del programa) por el ordenador.
Registro físico: Se denomina al conjunto de datos transferidos en una operación de lectura/escritura. A nivel mas interno, seria el conjunto de bytes que se transfieren en una operación de lectura/escritura de la memoria principal al dispositivo de almacenamiento o a la inversa.
Registros expandidos: Son registros logicos muy largos que han de ser leidos en varios registros fisicos, debido a su tamaño.
Campo (field): Es cada uno de los diferentes datos que constituyen un registro lógico.
Clave (key): Es el campo del registro lógico que sirve para identificar al registro al que pertenece.
Se usa para localizar al registro dentro de un fichero y para ordenar el mismo.
Puede ser cualquier campo dentro del registro. No todos los ficheros tienen un campo clave.
Puede ser cualquier campo del registro y puede haber mas de un campo clave en un registro, denominado clave primera, secundaria, y así sucesivamente.
El numero de campos clave que puede tener un registro depende del lenguaje de programación que estemos usando.
Bloqueo de registros (factor de bloqueo): En general un registro físico puede constar de un numero variable de registros logicos, es decir, se pueden transferir varios registros lógicos de la memoria al soporte magnético o viceversa, empleando una sola operación de lectura y/o escritura. Esta operación recibe el nombre de bloqueo y los registros físicos así formados se llaman bloques. El numero de registros lógicos contenidos en un bloque recibe el nombre de factor de bloqueo.
Las ventajas del bloqueo de registros son:

- Mayor velocidad en los procesos de entrada y salida: Las operaciones de entrada y salida son las que consumen mayor tiempo en la ejecución de los programas al intervenir elementos mecánicos ajenos a la CPU. Este tiempo se reduce al aumentar al numero de bits que se transfieren en cada operación de lectura del disco a la memoria o de escritura de la memoria al disco.
- Mayor aprovechamiento de la capacidad del soporte de almacenamiento: Tanto en cintas magnéticas como en discos los bloques se graban separados por espacios interbloques. Estos espacios los utiliza el

sistema para realizar ciertas operaciones y guardar información. Obviamente, cuanto mayor sea el registro físico menos espacio interbloque existirá en el soporte.

Memorias intermedias (buffers): El sistema crea estos buffers dentro de la memoria principal y se utilizan para las operaciones de entrada y salida de programas y datos en el ordenador. El programador puede definir el numero de estas áreas si no fuesen suficientes las estandar del sistema.

Medidas de utilización de los archivos

ACTIVIDAD

Es el porcentaje de registros procesados en relación con el numero total de registros.

1. Tasa de actividad = nºreg procesados / nºreg totales * 100

VOLATILIDAD

Consiste en el porcentaje de registros que se adicionan, suprimen, o modifican en un fichero respecto al numero medio de registros del fichero (en un periodo de tiempo concreto).

Pues; se dice que un fichero es:

Volátil: si tiene un porcentaje de adiciones y supresiones alto.

Estático: si tiene un porcentaje de adiciones y supresiones bajo.

Depende de 3 tasas

1. Tasa de adición = nºreg añadidos / nºreg totales * 100
2. Tasa de Supresión = nºreg eliminados / nºreg totales * 100
3. Tasa de modificación = nºreg modificados / nºreg totales * 100
4. Tasa de crecimiento = Tasa de adición – Tasa de supresión

Memorias

Debemos también prestar atención a los tipos de memorias que hay en un sistema informático:

1. Memoria primaria: RAM y Cache. Memoria volátil y de acceso rápido para el entorno y gestión.
2. Memoria secundaria: Discos, Cintas, CDs, DVDs, etc. Memoria permanente y de lento acceso (en comparación con RAM)
3. Memoria terciaria: Cluster, Sector. Son aquellas partes de un dispositivo de almacenamiento a las que no podemos acceder de forma directa. Como el cluster o sector en el caso del HD.

Registros

Los ficheros se guardan en discos y están organizados en unas unidades llamadas registros, donde cada registro esta formado por campos.

Los registros pueden ser de tres tipos:

- De longitud fija: Caracterizado por que siempre va a ocupar el mismo espacio en el disco, tenga o no información el registro. Tres tipos:
    1. Utilizando el mismo número de campos en cada registro, con iguales longitudes de los campos componentes dentro de cada registro.
    2. Con igual número de campos componentes, con distinta longitud de cada campo dentro de cada registro.
    3. Con distinto número de campos en cada registro.

- De longitud indefinida: Un registro lógico formado por varios campos de tamaño variables. En los ficheros de este tipo el ordenador desconoce el tamaño (indefinido) de sus registros, y por tanto no puede acceder a la información (registro) directamente, debido a que al no saber el tamaño tampoco puede calcular la posición. El sistema de acceso a este tipo de registros es recorriendo secuencialmente los que le preceden. Los registros de este tipo contiene la siguiente información:
    1. El primer campo del registro al que se accede
    2. El último campo del registro al que se accede
    3. Un sólo campo del registro del que se accede

- De longitud variable: Pueden contener cualquier tamaño en bytes, se puede especificar previamente un máximo y un mínimo. Y el tamaño del registro oscila entre el máximo y el mínimo.
  Este tipo de registros se usaba mucho anteriormente, pero causaba ciertos problemas.
  Se usan unos métodos para poder predefinir la longitud de los registros con el fin de poder acceder a ellos de forma correcta y sin posibles errores.
    Separadores de campos (banderas): Se sitúa al inicio y final del campo un carácter especial y único que identifique el principio y el final del campo. Este carácter especial no se puede dar dentro del

propio contenido de los campos. El carácter elegido será el usado siempre para esa función en todo el registro y fichero.
Indicadores de longitud: Se sitúa al inicio y final del campo un campo auxiliar que almacena el tamaño de cada campo, con el fin de identificar su tamaño y por tanto su dimensión.
Máscaras: la ausencia o presencia de campos se indica en el primer campo del registro, utilizando subcampos conteniendo cero o uno según exista o no, el segundo, tercero, etc. campo del registro.

Procesamiento de ficheros o Acceso a registros

Es la forma usada para sacar la información (registros) de los ficheros que se encuentran almacenados en el soporte (cintas, discos, disco duro "HD", etc).
El tipo de soporte condiciona este acceso, así pues podemos distinguir entre:

1. Acceso secuencial: En este acceso los registros se leen uno a uno desde el registro primero hasta el registro que se busca, o hasta el final (si no se ha encontrado). Se puede usar tanto en dispositivos secuénciales como direccionales.

2. Acceso directo: Permite seleccionar un registro directamente (con un numero mínimo de lecturas) a trabes de su clave sin necesidad de buscar en ninguno mas. Este tipo de acceso puede realizarse de dos formas:

    - Cálculo: Cada registro viene con una clave implementada, sobre la que se aplica un cálculo (hashing) y el resultado de este ya indica el lugar de grabación (la dirección en memoria dentro del soporte).

    - Índice: Existe un index/índice asociado o independiente al fichero en el cual se busca el registro y nos dice en que dirección de memoria se encuentra dicho registro requerido.

    Hashing: Se trata de unos algoritmos ya realizados por programadores que obtienen nueceros aleatorios, pero siempre dentro del rango de capacidad del soporte usado. Esos números luego definen la dirección en memoria donde se va a almacenar el registro. El algoritmo hashing sea cual sea debe cumplir las siguientes condiciones
    - Maximizar el espacio disponible en el dispositivo de almacenamiento. Debe de dar como resultado prácticamente todos las direcciones posibles, con un margen muy amplio, si hay un margen grande de direcciones que nunca van a salir será peor, y así reduciremos el espacio en el dispositivo de almacenamiento.
    - Establecer una relación lógica entre la dirección física y la dirección lógica. O lo que es lo mismo una relación entre la clave obtenida o resultado (el que se guarda junto al registro en el fichero) y la dirección que contiene el registro.
    - Producir el menor numero de registros que con distintas claves nos creen las mismas direcciones de almacenamiento.
    - Que el abanico de resultados del hashing no se salga de las posibilidades de capacidad del dispositivo. Por ejemplo, tenemos un disco duro y supongamos que tenemos en el hasta 10.000 direcciones de memoria, en cada dirección de memoria podemos almacenar un dato. Pues el hashing debe devolver un valor menor o igual a 10.000, no puede dar un valor mayor, puesto que ese registro luego no se podría guardar en el disco.

        Algunas técnicas empleadas en hashing son: el truncamiento, extracción, selección, etc.

3. Acceso indexado: En este tipo de acceso se usa una tabla auxiliar que contiene la clave y la dirección relativa del registro que queremos seleccionar. Una vez localizado en esa tabla se accede directamente al registro.

4. Acceso dinámico: Se basa en un acceso directo a un registro y a los demás se accede secuencialmente. Va directo a unas marcas, luego de marca a marca va secuencial.

Organización de ficheros

La organización de los ficheros, es la forma de estructurar y almacenar datos en un dispositivo de almacenamiento.

Soportes: Son los dispositivos que almacenan los datos, existen dos tipos de soportes.

1. Soportes. Secuénciales o de acceso secuencial: Se usan principalmente para copias de seguridad, y también por razones de antigüedad, Ejpl: cintas magnéticas.
2. Soportes. Direccionales o de acceso directo: Son los de uso generalizado, los mas empleados, Ejpl: discos.

El tipo de organización de un fichero depende del dispositivo (soporte) en el que se va a almacenar.

1. Secuencial: Se almacena un registro detrás de otro y todos seguidos, sin orden.
2. Directa: Los registros se almacenan en función de la respuesta de un algoritmo de cálculo (hashing).
3. Indexada: Se almacenan secuencialmente, y acompañados por un índice, así que disponen de orden.

Organización o modo secuencial: En este modo los registros se disponen uno a continuación del otro. Este tipo de gestión puede usar dos métodos.

Simple: Uno detrás de otro, sin dejar huecos en blanco entre ellos.

- Ventajas
  1. Consultas muy rápidas para procesamiento secuencial, una vez que llegas a un bloque de registros el procesamiento de todos los registros que están en ese bloque se producen de forma secuencial y muy rápidamente.
  2. Ahorramos espacio en el soporte, puesto que al meter un registro va inmediatamente después del anterior, con lo cual no perdemos nada de espacio en soporte.
  3. Este modo podemos usarlo en cualquier tipo soporte.

- Desventajas:

  1. Para acceder al registro numero “n” en el fichero hay que recorrer primero “n-1” registros. De forma secuencial, con lo que retardamos la velocidad de proceso. Así pues deducimos que es lento para consultas puntuales.
  2. Para actualizar un registro, ya sea: añadir, eliminar, modificar. Debemos hacer una copia del fichero, debido a que a la hora de eliminar los registros se deben correr hacia atrás y el mismo problema con las otras acciones. A la hora de añadir solo puede ser al final, justo después del ultimo registro metido.
  3. Los registros de este método se encuentran almacenados de forma desordenada debido a que cada registro se mete a continuación del anterior.

Encadenadas: Son los ficheros que usan una organización secuencial pero ordenada por punteros, con lo que mejoran a los “simples” al estar ordenados. Los registros se procesan en orden lógico uno tras otro, pero su orden físico no tiene porque ser así (determinado por punteros).
Los registros de este modo disponen de un campo mas donde se almacena un puntero al registro anterior o siguiente.
Este tipo de organización se podemos observar su uso para algunos algoritmos propios de la metodología de la programación informática, como listas, listas múltiples, anillos, árboles.

- Ventajas/Desventajas: Las mismas que el metodo “simple”, con la gran diferencia de que en este lugar los registros si se encuentran ordenados. Una ventaja mas y una desventaja menos con respecto al anterior caso.

Representación grafica

En el ejemplo vemos como en la organización secuencial encadenada se situan los registros de forma ordenada gracias al nuevo campo de almacenamiento para punteros.

Caso 1

En este caso tenemos el fichero con 3 registros en la primera columna se muestra el numero de cada registro insertado (secuencialmente), en la segunda tenemos el dato del registro, y en la tercera el campo donde se almacena el puntero. El primer registro (A) vemos como tiene un puntero que apunta al registro 2 (C) y este a su vez al registro 3 (D).

Nº Registro	Registro (dato)	Puntero
1	A	2
2	C	3
3	D	1

Nº Registro	Registro (dato)	Puntero
1	A	4

Caso 2

En este ejemplo vamos a insertar un nuevo registro el numero 4, vemos como el registro 1 (A) pasa a apuntar al registro 4 (B) y este al registro 2 (C). Y así sucesivamente.

2	C	3
3	D	1

4	B	2

Organización o modo indexado: En este modo los registros pueden ser localizados a trabes de una tabla aparte llamada index o índice que contiene la dirección de cada uno de los registros que se encuentran en el fichero. Por lo tanto su función es acceder directamente a un registro basando su búsqueda en ese índice.
Este índice o tabla index a su vez se puede organizar de varias formas: secuencial, multi-nivel, árbol. A trabes de este índice podemos procesar el fichero de forma secuencial o de forma directa, será una u otra forma en función de la organización del índice, independientemente de cómo se encuentre organizado el fichero que contiene los registros.

Este método divide el espacio del soporte en 3 zonas.

1. Área primaria o de datos: Es la zona donde esta el contenido ordenado ascendentemente por el valor de su clave, este área se encuentra dividida en segmentos y cada segmento contiene un numero “n” de registros.

2. Área de índices: En este área los registros están formados por 2 campos el primero contiene la clave del ultimo registro de cada segmento y el segundo contiene la dirección de memoria del comienzo de cada segmento.

3. Área overflow o desborde: En ella se insertan los registros que no han sido incluidos en el área primaria y que tienen ahora intermedios, para los registros insertados en dicha área.

El área primaria y el índice no se alteran después de ser creado el fichero, el overflow si, este va aumentando con todos los registros que son insertados.

Representación grafica

Área primaria o de datos

Dirección	Datos de reg	Nº registro
1	*	2
2	*	30
3	*	40
4	*	60
5	*	90
6	*	120
7	*	180

Reg	Dato
42	*
98	*

Área de desborde

Área index (índice)

Sector	Reg
30	1
60	3
120	5

Primera grafico (area primaria o de datos): Se origina al crear el fichero y no se modifica nunca ni se altara con la inserción de nuevos registros y demas operaciones relacionadas con los registros.
En este area, la primera columna esta ocupada por el campo que almacena la direccion de memoria donde se encuentra guardado el registro.
En estos ejemplos puse numeros enteros, para que su comprensión sea mas facil, pero en realidad serian hexadecimales.
Y la tercera columna guarda el numero del registro, suponemos que entree el registro 2 y el 30 hay otros (15, 17, 24, etc).

Las flechas indican la situación de cada sector, en este ejemplo tenemos 3 sectores, en el registro 30, 60 y 90.

Segundo grafico (area index): Es el tipo de indice usado por este metodo.
La primera columna almacena el registro donde se hala el sector.
La segunda columna almacena el numero del registro.

Tercer grafico (area de desborde): Es donde se almacenan todos los registros que se van a insertar.
La primera columna almacena el numero del registro.
La segunda columna almacena el dato.

Explicación: Cuando insertamos un registro a este fichero va a parar a la zona de overflow o desborde, como el 42 y 98 que están ahora. Ahora bien, cuando deseamos realizar una operación con esos dos registros, lo que hace este método es, primero coge el numero del registro, por ejemplo "42", luego teniendo ese numero acude a la tabla de índice (indexado) en ella se va preguntando, ¿es 42 mayor que 1 y menor que 30?, y así sucesivamente hasta que la condicional es afirmativa (true). En el caso del 42 daría verdadero en la segunda fila del index. Una vez que da verdadero pasa a buscar en el área primaria el registro desde el sector del registro 3 hasta el registro 60. Si el numero se encuentra lo da y si no pasa al área de overflow a buscarlo allí. En el área del overflow se busca de forma secuencial, con lo cual cuanto mas grande el overflow mas retardara y será peor el método.

Modo indexado secuencial encadenado: Este método usa un fichero de datos secuencial y un índice con punteros. Con lo que aprovecha lo mejor de los ficheros encadenados e indexados, esto es: usa punteros o índices. Este método es igual que el anterior pero añade la ventaja de los punteros, con lo que encadena así los registros.

Organización por agrupamiento o clustering: En este tipo de almacenamiento, se agrupan tablas cuyos ficheros comparten algunos atributos (campos), a los que se llama claves de agrupamiento.
Ejemplo: Tenemos un fichero y en el se almacenan unas tablas (siguiendo este método), cada tabla almacena a su vez a varios equipos de hokey y dentro de esos equipos se hallan los jugadores de cada uno.

Ventajas
El acceso a equipos es rápido y a los jugadores del equipo también es rápido.

Desventaja
Si se busca un jugador en concreto resulta complejo y bastante lento.

Eliminando datos del fichero

Para borrar datos o registros de un fichero se pueden usar las siguientes técnicas:
Marcando los registros para indicar que estan borrados.
Cambiando la direccion de los punteros.
Eliminando el registro fisicamente

Para que la zona de los elementos borrados no crezca mucho se usan unas técnicas para usar ese espacio, que son:

1. Lista de registros disponibles: Consiste en disponer de una lista que contenga toda las direcciones de los registros eliminados. Esta tecnica es una mejora de otra que consiste en recorrer el fichero hasta que encuentra el primer hueco vacio.
2. Lista de registros de longitud fija: En este se crea una lista pero con la posición del registro dentro del fichero.
3. Lista de registros de longitud variable: Consiste en crear una lista donde aproveche las direcciones fisicas de inicio de cada registro y la longitud del registro.

Fragmentación y compresión (compactacion de datos)
Un fichero esta fragmentado cuando esta desaprovechando espacio.
Se puede distinguir 2 tipos defragmentacion:

- Fragmentacion interna: se produce cuando el espacio no usado esta asignado a un registro.
- Fragmentacion externa: consiste en una utilización inadecuada del disco o soporte.

Las tecnicas o estrategias de colocacion que se aplican para reducir espacios en registros de longitud variable.
Primer ajuste: Primer hueco que encuentra vacio, ahí va.
Mejor ajuste: Consiste en ordenar ascendentemente los espacios vacios para optimizar la búsqueda. Funciona mejor con fragmentacion interna.
Peor ajuste: Los mismo que la anterior pero la lista que crea es de orden descendente. Y funciona mejor con la fragmentación externa.

Compresión: Se usa para que ocupe menos espacio los datos dentro del disco.
Existen 2 tipos de compresión
Compresión irreversible: Esta técnica es usada para imágenes, voz (mp3). Cuando se produce la compresión y se desea volver a la calida anterior