MÓDULO 6 | ARCHIVOS - CONCEPTOS BASICOS

1. CONCEPTOS BASICOS

El almacenamiento de datos en variables y vectores es temporal; los datos se pierden cuando una variable

sale de su ámbito o alcance de influencia, o bien cuando se termina el programa.

La mayoría de las aplicaciones requieren que la información se almacene de forma persistente, es decir que no se borre o elimine cuando se termina la ejecución del programa. Por otra parte, en numerosas aplicaciones se requiere utilizar grandes cantidades de información que, normalmente, no caben en la memoria principal.

Debido a estas causas se requiere utilizar archivos (ficheros) para almacenar de modo permanente grandes cantidades de datos, incluso después que los programas que crean los datos se terminan. Estos datos almacenados en archivos se conocen como datos persistentes y permanecen después de la duración de la ejecución del programa.

Las computadoras almacenan los archivos en dispositivos de almacenamiento secundarios, tales como discos CD,

DVD, memorias flash USB, memorias de cámaras digitales, etc.

El procesamiento de archivos es una de las características más importantes que un lenguaje de programación debe tener para soportar aplicaciones comerciales que procesan, normalmente, cantidades masivas de datos persistentes.

La entrada de datos normalmente se realiza a través del teclado y la salida o resultados van a la pantalla. Estas operaciones, conocidas como Entrada/Salida (E/S), se realizan también hacia y desde los archivos.

La noción de archivo o fichero está relacionada con los conceptos de:

• Almacenamiento permanente de datos.

• Fraccionamiento o partición de grandes volúmenes de información en unidades más pequeñas que puedan ser almacenadas en memoria central y procesadas por un programa.

Un **archivo** o fichero
es un conjunto de **datos** estructurados
en una colección de entidades elementales o básicas denominadas **registros** que son de igual tipo y constan a su vez de diferentes entidades de nivel más bajo denominadas **campos.**

Veamos ahora cada uno de estos conceptos básicos:

1.1. DATOS

El *dato* es una representación simbólica (número, símbolo), una *característica de una entidad* , que pueden describir o explicar hechos empíricos, acontecimientos, y a las entidades mismas que se están estudiando. Un dato aislado puede no ser importante para el objeto de estudio que nos ocupa. Sólo cuando un conjunto de datos se examina conjuntamente al significado, se puede apreciar la información de forma coherente contenida en ellos. Los datos agrupados, estructurados e interpretados son la base de la información relevante que podemos procesar sistemáticamente.

1.2. CAMPOS

Un *campo* es un *item* o *elemento de datos elementales*, tales como un nombre, número de empleados, ciudad, número de identificación, etc. Un campo está caracterizado por su tamaño o longitud y su tipo de datos (cadena de caracteres, entero, lógico, etcétera.). Los campos pueden incluso variar en longitud. En la mayoría de los lenguajes de programación los campos de longitud variable no están soportados y se suponen de longitud fija.

Campos

Nombre	Dirección	Fecha de nacimiento	Estudios	Salario	Trienios

Campos de un registro

Un campo es la unidad mínima de información de un registro.
Los datos contenidos en un campo se dividen con frecuencia en *subcampos*; por ejemplo, el campo fecha se divide en los subcampos día, mes, año.

Campo	0	7	0	7	1	9	9	5
Subcampo	Día		Mes		Año			

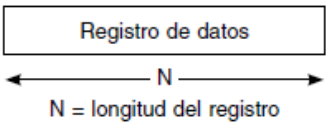
Los rangos numéricos de variación de los subcampos anteriores son:
1 ≤ dia ≤ 31
1 ≤ mes ≤ 12
1 ≤ año ≤ 1987

1.3. REGISTROS

Un *registro* es una colección de información, normalmente relativa a una entidad particular. Un registro es una colección de campos lógicamente relacionados, que pueden ser tratados como una unidad por algún programa. Un ejemplo de un registro puede ser la información de un determinado empleado que contiene los campos de nombre, dirección, fecha de nacimiento, estudios, salario, trienios, etc.

Los registros pueden ser todos de *longitud fija*; por ejemplo, los registros de empleados pueden contener el mismo número de campos, cada uno de la misma longitud para nombre, dirección, fecha, etc. También pueden ser de *longitud variables*.

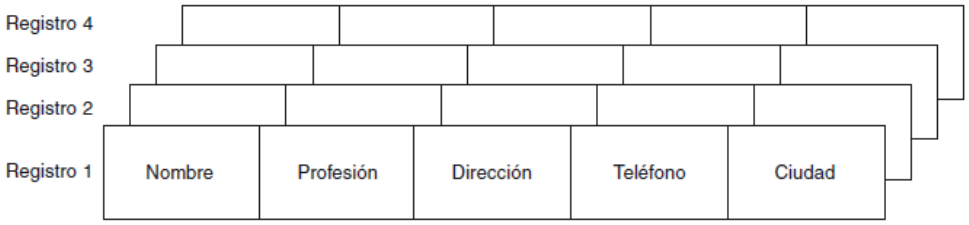
Los registros organizados en campos se denominan *registros lógicos*.



1.4. ARCHIVOS

Un *fichero* (*archivo*) de datos —o simplemente un **archivo**— es una colección de registros relacionados entre sí con aspectos en común y organizados para un propósito específico. Por ejemplo, un fichero de una clase escolar contiene un conjunto de registros de los estudiantes de esa clase. Otros ejemplos pueden ser el fichero de nóminas de una empresa, inventarios, stocks, etc.

La Figura siguiente recoge la estructura de un archivo correspondiente a los suscriptores de una revista de informática.



Estructuras de un archivo "suscriptores".

Un archivo en una computadora es una estructura diseñada para contener datos. Los datos están organizados de tal modo que puedan ser recuperados fácilmente, actualizados o borrados y almacenados de nuevo en el archivo con todos los campos realizados.

1.5. BASES DE DATOS

Una colección de archivos a los que puede accederse por un conjunto de programas y que contienen todos ellos datos relacionados constituye una base de datos. Así, una base de datos de una universidad puede contener archivos de estudiantes, archivos de nóminas, inventarios de equipos, etc.

1.6. ESTRUCTURA JERÁRQUICA

Los conceptos carácter, campos, registro, archivo y base de datos son *conceptos lógicos* que se refieren al medio en que el usuario de computadoras ve los datos y se organizan. Las estructuras de datos se organizan de un modo jerárquico, de modo que el nivel más alto lo constituye la base de datos y el nivel más bajo el carácter.

1.7. JERARQUÍA DE DATOS

Una computadora, procesa todos los datos como combinaciones de ceros y unos. Tal elemento de los datos se denomina bit (binary digit). Sin embargo, como se puede deducir fácilmente, es difícil para los programadores trabajar con datos en estos formatos de bits de bajo nivel. En su lugar, los programadores prefieren trabajar con caracteres tales como los dígitos decimales (0-9), letras (A-Z y a-z) o símbolos especiales (&, *, , @, €, #,...). El conjunto de todos los caracteres utilizados para escribir los programas se denomina *conjunto* o *juegos de caracteres* de la computadora.

Al igual que los caracteres se componen de bits, los *campos* se componen de caracteres o bytes. Un *campo* es un grupo de caracteres o bytes que representan un significado. Por ejemplo, un campo puede constar de letras mayúsculas y minúsculas que representan el nombre de una ciudad.

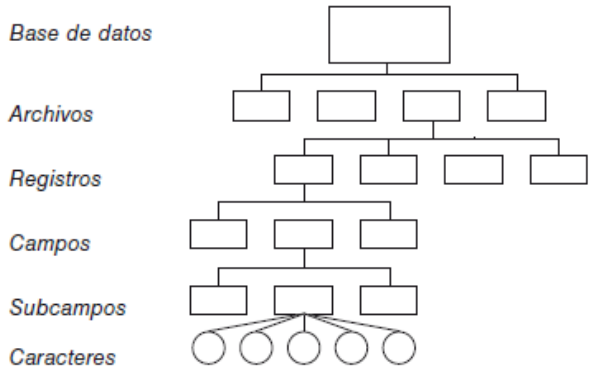
Los datos procesados por las computadoras se organizan en *jerarquías de datos* formando estructuras a partir de bits, caracteres, campos, etc.

Los campos se agrupan en *registros* . Un registro es un grupo de campos relacionados que se implementan con tipos de datos básicos o estructurados. En un sistema de matrícula en una universidad, un registro de un alumno o de un profesor puede constar de los siguientes campos:

- Nombre (cadena).
- Número de expediente (entero).
- Número de Documento Nacional de Identidad o Pasaporte (entero doble).
- Año de nacimiento (entero).
- Estudios (cadena).

Un archivo es un grupo de registros relacionados. Así, una universidad puede tener muchos alumnos y profesores, y un archivo de alumnos contiene un registro para cada empleado. Un archivo de una universidad puede contener miles de registros y millones o incluso miles de millones de caracteres de información.

Las Figura siguiente muestra las estructuras jerárquicas de datos.



Los registros poseen una *clave* o llave que identifica a cada registro y que es única para diferenciarla de otros registros. En registros de nombres es usual que el campo clave sea el pasaporte o el DNI (Documento Nacional de Identidad).

Un conjunto de archivos relacionados se denomina base de datos. En los negocios o en la administración, los datos se almacenan en bases de datos y en muchos archivos diferentes. Por ejemplo, las universidades pueden tener archivos de profesores, archivos de estudiantes, archivos de planes de estudio, archivos de nóminas de profesores y de PAS (Personal de Administración y Servicios). Otra jerarquía de datos son los

sistemas de gestión de bases de datos (**SGBD** o **DBMS**) que es un conjunto de programas diseñados para crear y administrar bases de datos.

RESUMIENDO

Un **campo** es un conjunto de caracteres o bytes que representan un significado, y se agrupan en registros.

Un **registro** es un grupo de campos relacionados que se implementan con tipos de datos básicos o estructurados , y se agrupan en archivos. Cada registro posee una clave que lo hace único e irrepetible.

Un **archivo** es un grupo de registros relacionados

Un conjunto de archivos relacionados se denomina **base de datos**

2. SOPORTES SECUENCIALES Y DIRECCIONABLES

El soporte es el medio físico donde se almacenan los datos. Los tipos de soporte utilizados en la gestión de archivos son:

• *Soportes secuenciales.* Son aquellos en los que los registros —informaciones— están escritos unos a continuación de otros y para acceder a un determinado registro *n* se necesita pasar por los *n – 1* registros anteriores.

• *Soportes direccionables.* Se estructuran de modo que las informaciones registradas se pueden localizar directamente por su dirección y no se requiere pasar por los registros precedentes. En estos soportes los registros deben poseer un campo clave que los diferencie del resto de los registros del archivo. Una dirección en un soporte direccionable puede ser número de pista y número de sector en un disco. Los soportes direccionables son los discos magnéticos, aunque pueden actuar como soporte secuencial.

3. ORGANIZACIÓN DE ARCHIVOS

Según las características del soporte empleado y el modo en que se han organizado los registros, se consideran dos tipos de acceso a los registros de un archivo:

• *Acceso secuencial* implica el acceso a un archivo según el orden de almacenamiento de sus registros, uno tras otro.

• *Acceso directo* implica el acceso a un registro determinado, sin que ello implique la consulta de los registros precedentes. Este tipo de acceso sólo es posible con soportes direccionables.

La *organización* de un archivo define la forma en la que los registros se disponen sobre el soporte de almacenamiento, o también se define la organización como la forma en que se estructuran los datos en un archivo. En general, se consideran tres organizaciones fundamentales:

• *Organización secuencial.*

• *Organización directa o aleatoria ("random").*

• *Organización secuencial indexada ("indexed").*

3.1. Organización secuencial

Un archivo con organización secuencial es una sucesión de registros almacenados consecutivamente sobre el soporte externo, de tal modo que para acceder a un registro *n* dado es obligatorio pasar por todos los *n – 1* artículos que le preceden.

Los registros se graban consecutivamente cuando el archivo se crea y se debe acceder consecutivamente cuando se leen dichos registros.

- El orden físico en que fueron grabados (escritos) los registros es el orden de lectura de los mismos.
- Todos los tipos de dispositivos de memoria auxiliar soportan la organización secuencial.

Los archivos organizados secuencialmente contienen un registro particular —el último— que contiene una marca fin de archivo (**EOF** o bien **FF**). Esta marca fin de archivo puede ser un carácter especial como **'*'**.

3.2. Organización directa

Un archivo está organizado en modo directo cuando el orden físico no se corresponde con el orden lógico. Los datos se sitúan en el archivo y se accede a ellos directamente mediante su posición, es decir, el lugar relativo que ocupan.

Esta organización tiene la *ventaja* de que se pueden leer y escribir registros en cualquier orden y posición. Son muy rápidos de acceso a la información que contienen.

La organización directa tiene el *inconveniente* de que necesita programar la relación existente entre el contenido de un registro y la posición que ocupa. El acceso a los registros en modo directo implica la posible existencia de huecos libres dentro del soporte y, por consecuencia, pueden existir huecos libres entre registros.

La correspondencia entre clave y dirección debe poder ser programada y la determinación de la relación entre el registro y su posición física se obtiene mediante una fórmula.

Las condiciones para que un archivo sea de organización directa son:

- Almacenado en un soporte direccionable.
- Los registros deben contener un campo específico denominado *clave* que identifica cada registro de modo único, es decir, dos registros distintos no pueden tener un mismo valor de clave.
- Existencia de una correspondencia entre los posibles valores de la clave y las direcciones disponibles sobre el soporte.

Un soporte direccionable es normalmente un disco o paquete de discos.

3.3. Organización secuencial indexada

Un diccionario es un archivo secuencial, cuyos registros son las entradas y cuyas claves son las palabras definidas por las entradas. Para buscar una palabra (una clave) no se busca secuencialmente desde la “a” hasta la “z”, sino que se abre el diccionario por la letra inicial de la palabra. Si se desea buscar “índice”, se abre el índice por la letra *I* y en su primera página se busca la cabecera de página hasta encontrar la página más próxima a la palabra, buscando a continuación palabra a palabra hasta encontrar “índice”. El diccionario es un ejemplo típico de archivo secuencial indexado con dos niveles de índices, el nivel superior para las letras iniciales y el nivel menor para las cabeceras de página. En una organización de computadora las letras y las cabeceras de páginas se guardarán en un archivo de índice independiente de las entradas del diccionario (archivo de datos). Por consiguiente, cada archivo secuencial indexado consta de un archivo índice y un archivo de datos.

Un archivo está organizado en forma secuencial indexada si:

• El tipo de sus registros contiene un campo clave identificador.

• Los registros están situados en un soporte direccionable por el orden de los valores indicados por la clave.

• Un índice para cada posición direccionable, la dirección de la posición y el valor de la clave; en esencia, el índice contiene la clave del último registro y la dirección de acceso al primer registro del bloque.

Un archivo en organización secuencial indexada consta de las siguientes partes:

• *Área de datos o primaria*: contiene los registros en forma secuencial y está organizada en secuencia de claves

sin dejar huecos intercalados.

• *Área de índices*: es una tabla que contiene los niveles de índice, la existencia de varios índices enlazados se denomina *nivel de indexación*.

• *Área de desbordamiento o excedentes*: utilizada, si fuese necesario, para las actualizaciones.

El área de índices es equivalente, en su función, al índice de un libro. En ella se refleja el valor de la clave identificativa más alta de cada grupo de registros del archivo y la dirección de almacenamiento del grupo.

Los archivos secuenciales indexados presentan las siguientes *ventajas*:

• Rápido acceso.

• El sistema de gestión de archivos se encarga de relacionar la posición de cada registro con su contenido mediante la tabla de índices.

Y los siguientes *inconvenientes*:

• Desaprovechamiento del espacio por quedar huecos intermedios cada vez que se actualiza el archivo.

• Se necesita espacio adicional para el área de índices.

Los soportes que se utilizan para esta organización son los que permiten el acceso directo —los discos magnéticos—. Los soportes de acceso secuencial no pueden utilizarse, ya que no disponen de direcciones para las posiciones de almacenamiento.

4. OPERACIONES SOBRE ARCHIVOS

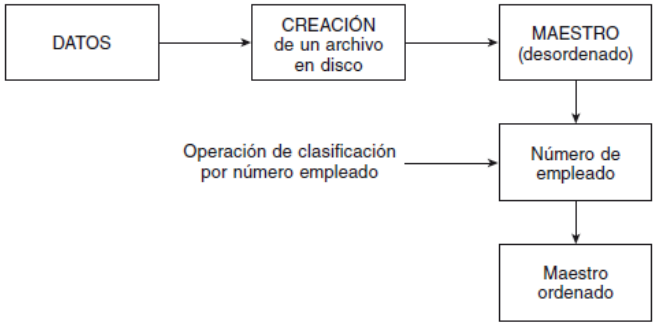
Tras la decisión del tipo de organización que ha de tener el archivo y los métodos de acceso que se van a aplicar para su manipulación, es preciso considerar todas las posibles operaciones que conciernen a los registros de un archivo. Las distintas operaciones que se pueden realizar son:

- *Creación.*
- *Consulta.*
- *Actualización* (altas, bajas, modificación, consulta).
- *Reorganización.*
- *Destrucción* (borrado).
- *Reunión, fusión.(merge)*
- *Rotura, estallido. (split)*

4.1. Creación de un archivo

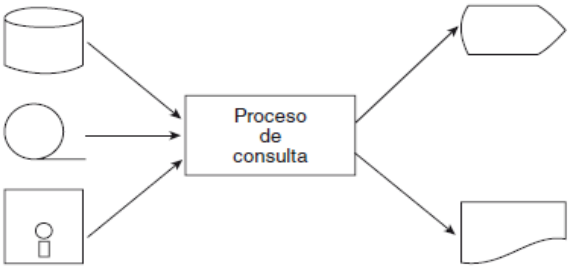
Es la primera operación que sufrirá el archivo de datos. Implica la elección de un entorno descriptivo que permita un ágil, rápido y eficaz tratamiento del archivo.
Para utilizar un archivo, éste tiene que existir, es decir, las informaciones de este archivo tienen que haber sido almacenadas sobre un soporte y ser utilizables. La *creación* exige organización, estructura, localización

o reserva de espacio en el soporte de almacenamiento, transferencia del archivo del soporte antiguo al nuevo.
Un archivo puede ser creado por primera vez en un soporte, proceder de otro previamente existente en el mismo o diferente soporte, ser el resultado de un cálculo o ambas cosas a la vez.
La Figura siguiente muestra un organigrama de la creación de un archivo ordenado de empleados de una empresa por el campo clave (número o código de empleado).



4.2. Consulta de un archivo

Es la operación que permite al usuario acceder al archivo de datos para conocer el contenido de uno, varios o todos los registros.

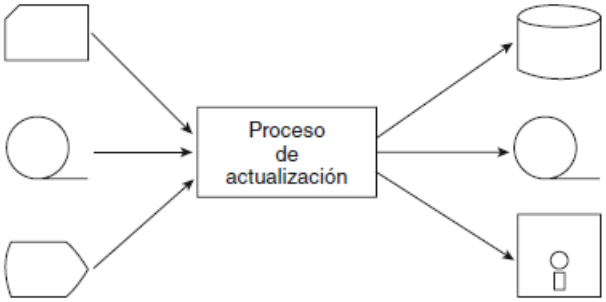


4.3. Actualización de un archivo

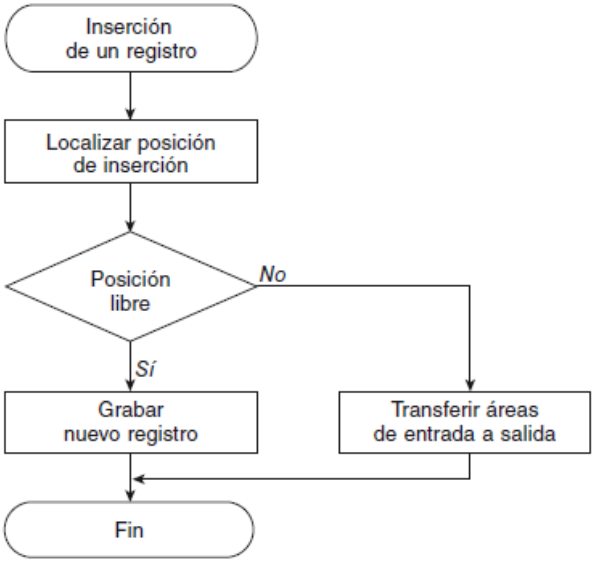
Es la operación que permite tener actualizado (puesto al día) el archivo, de tal modo que sea posible realizar las siguientes operaciones con sus registros:

- *Consulta* del contenido de un registro.
- *Inserción* de un registro nuevo en el archivo.
- *Supresión* de un registro existente.
- *Modificación* de un registro.

Un ejemplo de actualización es el de un archivo de un almacén, cuyos registros contienen las existencias de cada artículo, precios, proveedores, etc. Las existencias, precios, etc., varían continuamente y exigen una actualización simultánea del archivo con cada operación de consulta.



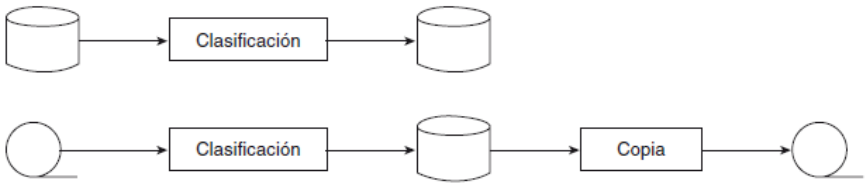
Actualización de un archivo (I).



Actualización de un archivo (II).

4.4. Clasificación de un archivo

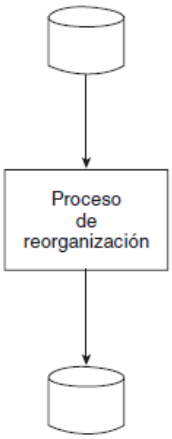
Una operación muy importante en un archivo es la *clasificación u ordenación* (*sort*, en inglés). Esta clasificación se realizará de acuerdo con el valor de un campo específico, pudiendo ser *ascendente* (creciente) o *descendente* (decreciente):



4.5. Reorganización de un archivo

Las operaciones sobre archivos modifican la estructura inicial o la óptima de un archivo. Los índices, enlaces (punteros),zonas de sinónimos, zonas de desbordamiento, etc., se modifican con el paso del tiempo, lo que hace a la operación de acceso al registro cada vez más lenta.

La reorganización suele consistir en la copia de un nuevo archivo a partir del archivo modificado, a fin de obtener una nueva estructura lo más óptima posible.

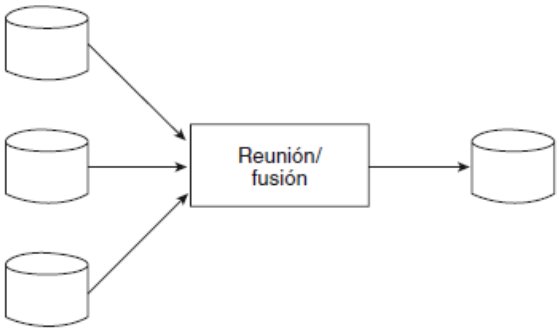


4.6. Destrucción de un archivo

Es la operación inversa a la creación de un archivo (*kill*, en inglés). Cuando se destruye (anula o borra) un archivo, éste ya no se puede utilizar y, por consiguiente, no se podrá acceder a ninguno de sus registros

4.7. Reunión, fusión de un archivo

Reunión. Esta operación permite obtener un archivo a partir de otros varios Se realiza una fusión cuando se reúnen varios archivos en uno solo, intercalándose unos en otros, siguiendo unos criterios determinados.



4.8. Rotura/estallido de un archivo

Es la operación de obtener varios archivos a partir de un mismo archivo inicial.