Organización de Archivos

Equipo docente

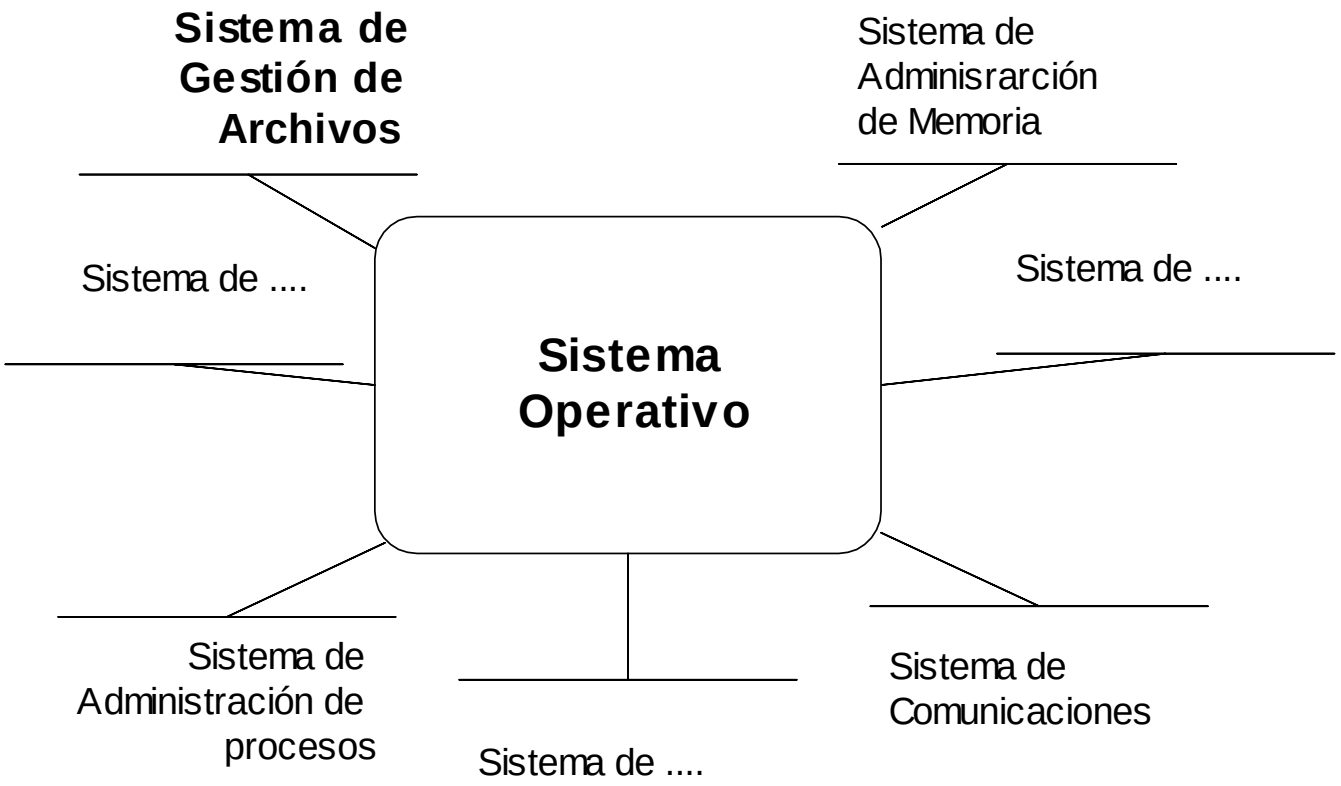
Prof. Adj. Lic. Carlos Rodriguez

J.T.P. Lic. Guillermo Cherencio

Bibliografía

Sistemas Operativos. W. Stallings, 4 ed. Prentice Hall. Cap. 12
Sistemas Operativos. A. Silberschatz. 5 ed. Pearson. Cap. 10,11,12
Fundamentos de Sistemas de Base de Dato, Elmasri, Navathe. Addison Wesley cap. 5 y 6
Introducción a las Bases de Datos, Mendelzon y Ale. Pearson. 2000. Cap. 6.

Sistema de Gestión de Archivos



Sistema de Gestión de Archivos

Provee mecanismos destinados al almacenamiento y acceso a archivos de programas y de datos por parte de procesos y usuarios en línea.

Presenta una vista lógica uniforme del almacenamiento de la información, dado que el S.O abstrae las características físicas del hardware

Compuesto por:

Colección Archivos
+ Estructura de Directorios
+ Particiones

Sistema de Gestión de Archivos

<table>
<tr><td rowspan="2">Partición A</td><td>directorio</td><td rowspan="4">Disco 1</td></tr>
<tr><td>archivos</td></tr>
<tr><td rowspan="2">Partición B</td><td>directorio</td></tr>
<tr><td>archivos</td></tr>
<tr><td rowspan="2">Partición C</td><td>directorio</td><td>Disco 2</td></tr>
<tr><td>archivos</td><td>Disco 3</td></tr>
</table>

Sistema de Gestión de Archivos

Objetivos

Cumplir con los requerimientos de usuarios

Garantizar datos válidos (no corruptos)

Optimizar el rendimiento (optimización de accesos, fragmentación)

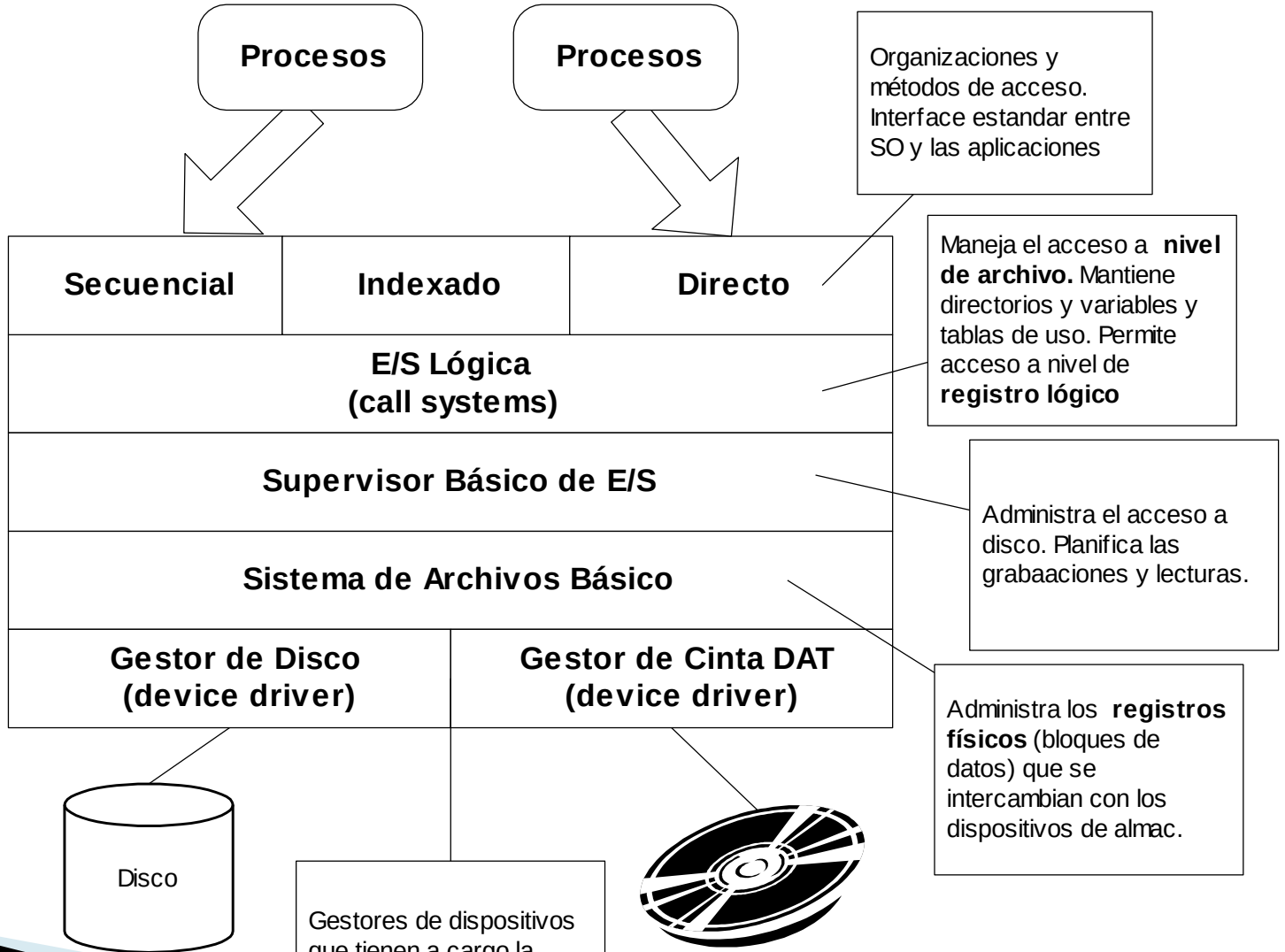
Soportar diversos dispositivos de almacenamiento

Minimizar pérdidas de datos (soportar respaldo, copias espejo)

Ofrecer a los usuarios y procesos una interface normalizada de E/S

Proporcionar soporte multiusuario.

Sistema de Gestión de Archivos



Archivos

"Unidad de almacenamiento lógica persistente, formada por una colección de información relacionada, grabada en una memoria secundaria bajo un nombre"

Tipos

- Datos
- Textos
- Programa

(todos pueden ser de texto o binarios)

La estructura de un archivo está definida por su tipo (jpg, exe, doc, etc)

Archivos

Tipo	Extensión	Función
Binario	exe, com, bin	prg en lenguaje de máquina.
Objeto	obj, o	Compilado pero no enlazado.
Fuente	c, pl, bas, pas, asm	prg fuente
datos	dat, txt	Archivo auxiliar de datos
Biblioteca	lib, a, pm	Librerias de rutinas
Multimedia	gif, jpg, wav, mp3, mpg	
Archivado (empaq), compresión	rar, tar, zip, arc	

Archivos

Atributos:

Nombre simbólico
Tipo
Ubicación (path)
Tamaño (en bytes, bloques)
Protección (para control de accesos) (rwx rwx rwx usr-id grp)
ID-Usuario dueño
Hora/fecha (creación, última modificación)

Terminología Relacionada con los Archivos de Datos

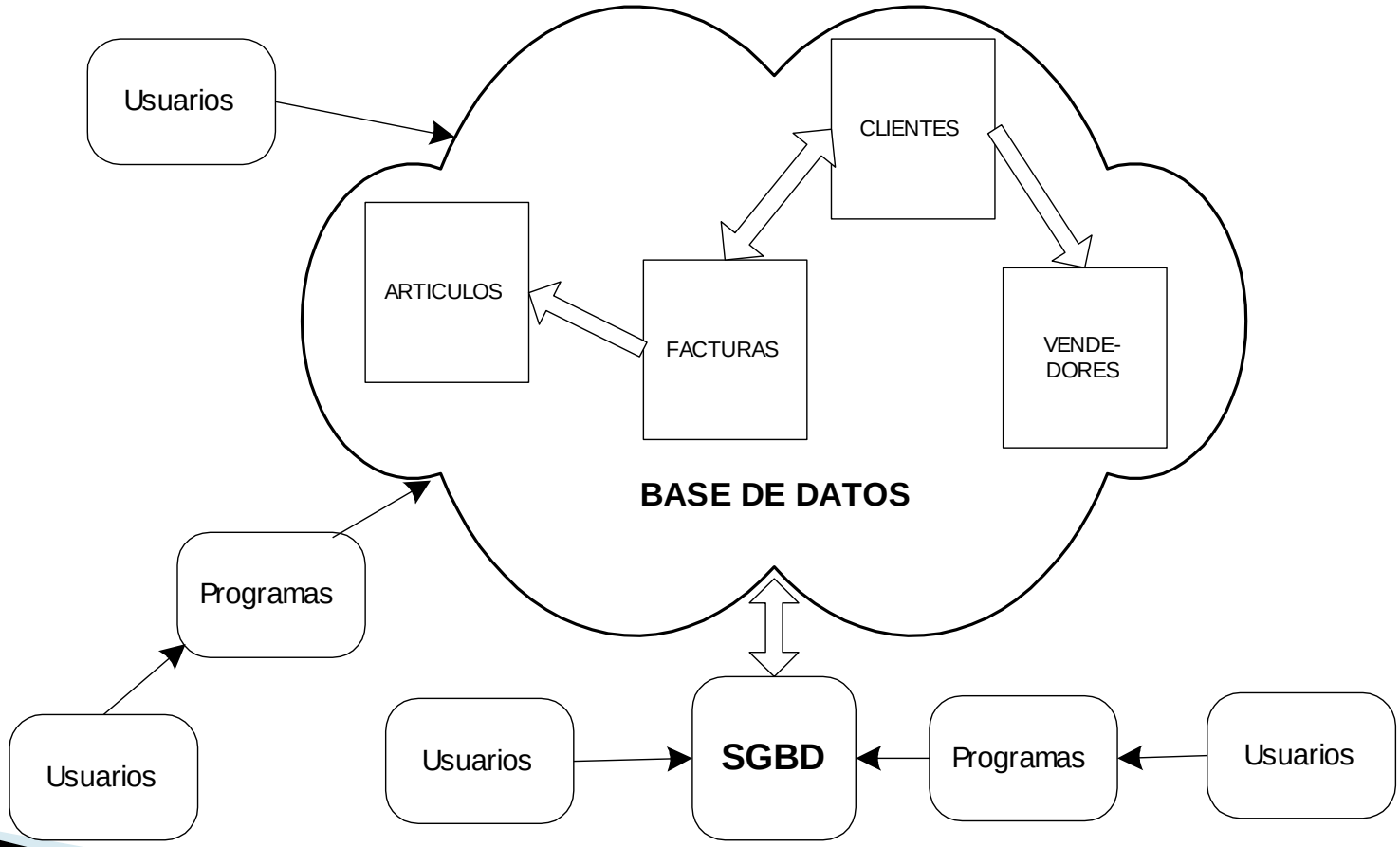
CAMPO: Son elementos básicos de datos de caracter atómico. Posee por atributos: un nombre, un tipo de datos, un dominio y un longitud.

REGISTRO LOGICO: Conjunto de campos relacionados que son tratados como una unidad por un programa de aplicación. Semánticamente son atributos que describen a una entidad.

ARCHIVO: Colección de registros del mismo tipo grabados en una memoria secundaria bajo un mismo nombre.

BASE DE DATOS: Conjunto de datos relacionados con propiedades.

Terminología Relacionada con los Archivos de Datos



Operaciones Básicas con Archivos

Las operaciones básicas o primitivas sobre archivos se implementan a través de llamadas al S.O. (call systems) proporcionadas por el subsistema de gestión de archivos a los programas de aplicación.

ABRIR	open(AP, "nombre") {append, output, read}
ESCRIBIR	write(AP,buffer) {a partir de donde está el puntero}
LEER	read(AP,buffer, tamaño)
REUBICARSE	seek(AP,posición)
CERRAR	close(AP)
TRUNCAR	Borrar el contenido de un archivo, pero mantiene sus metadatos (open(AP, ">pepe"))
CREAR	touch, open(AP,">pepe")
ELIMININAR	del, rm

Silberchatz presenta solo: crear, escribir, leer, seek, borrar y truncar.

Cómo el SO Identifica a los Archivos

UNIX Cada archivo es una secuencia de bytes. El sistema operativo no hace interpretación alguna.

System (Apple Macintosh)Cada archivo posee dos estructuras de datos asociadas: a) Resource fork (rama de recursos) almacena metadatos (creador, ícono, label, etc)
b) Data fork (rama de datos) es donde se almacenan los datos.

Microsoft Reconoce tipos por extensión.

Organizaciones **Básicas** de Archivos Vistas en Programación I

ORGANIZACION: Forma en la cual se organizan los registros de un archivo, sobre un soporte permanente.

A) SECUENCIAL: Registros de longitud variable
Utiliza delimitadores de campo y registro.
Solo soporta acceso secuencial
Soportada por todos los dispositivos
Se optimiza el uso del espacio

B) RELATIVA: Registros de longitud fija
No utiliza delimitadores
Soporta acceso secuencial y directo
Solo es soportada por dispositivos direccionables
Se optimiza el tiempo de acceso a los registros

Organizaciones **Básicas** de Archivos Vistas en Programación I

SECUENCIAL

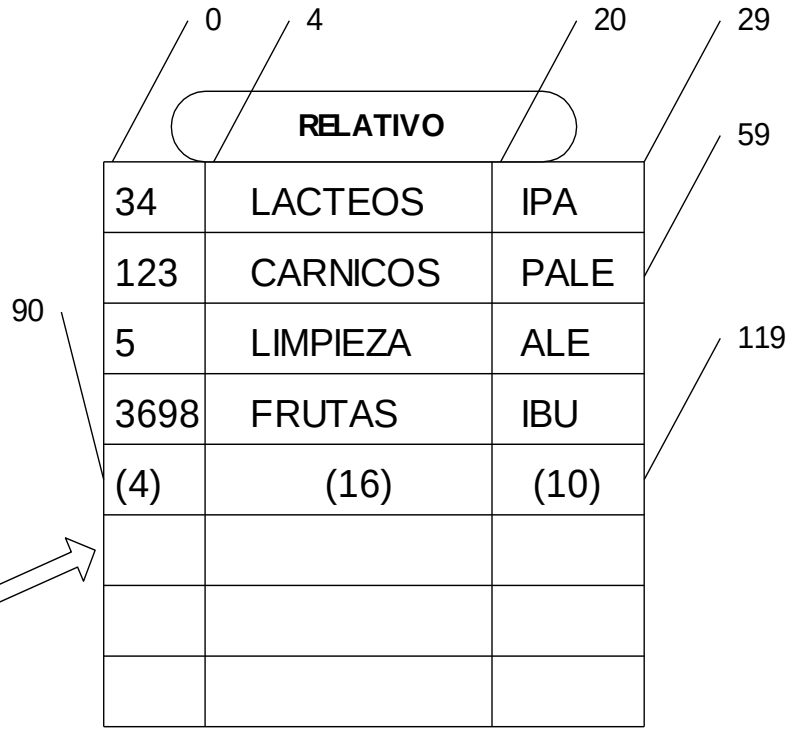
"Carlos",234,"masc","01/03/200"
"Ana",,"fem","12/11/1999"
"Silvia",1983,"fem","23/06/1970"

Calculo de la posición de inicio de un registro n

n = número de registro
l = largo del registro

Posición_Inicio(n) = (n * (l)) - l

Posición_Inicio(1) = (1 * (30)) - 30 = 0
Posición_Inicio(2) = (2 * (30)) - 30 = 30
Posición_Inicio(4) = (4 * (30)) - 30 = 90



Recuperación de Datos versus Recuperación de Información

	Recuperación de datos	recuperación de información
Match	Exacto (booleano)	Parcial (por distancia)
Modelo	Determinístico	Probabilístico
Lenguaje de consulta	Artificial (ej SQL)	Natural
Items buscados	Matching	Relevancia (métrica de importancia)
Respuesta al error	Sensitiva	No sensitiva

RECUPERANDO DATOS

Claves

CLAVE CANDIDATA: Conjunto de atributos que identifican UNIVOCA y MINIMAMENTE a un registro en un archivo. En caso de que no sean mínimas se las conoce como SUPERCLAVE

CLAVE PRIMARIA: Es una clave candidata, seleccionada por el diseñador como la principal para el acceso

Registro_Cliente(DNI, CUIL, ID_CLIENTE, NOMBRE, DIRECCION, CP)

CUIL, DNI, ID_CLIENTE son claves **candidatas**
ID_CLIENTE clave **primaria**

(CUIL,CP) es una **superclave** (*no es mínima*, pero identifica univocamente)

DOMINIO: Conjunto de instancias válidas $I(i_1,..i_n)$ para un atributo A

RECUPERANDO DATOS

Claves

CLAVE ALTERNATIVA: Son aquellas claves candidatas que no fueron seleccionadas como primarias.

CAMPO PRIMO: Son aquellas campos que **forman** parte de la clave candidata.

CLAVE EXTRANJERA: (FORANEA) Se utiliza a los efectos de relacionar dos archivos La clave primaria C2 de un archivo A2 se inserta en los registros de un archivo A1.

CLAVE SECUNDARIA: Pueden ser simples o compuestas, y admiten repetición de valores de clave. Se definen a los efectos de tener un acceso rápido a ciertos registros.

RECUPERANDO DATOS

Claves

INTEGRIDAD: Se refiere a la exactitud o corrección de los datos sobre un conjunto de archivos *afines*

La ***perdida accidental de la consistencia*** de los datos puede ser resultado de :

1. Caídas durante el procesamiento

2. Anomalías causadas por la distribución de datos entre varias computadoras

3. La modificación no autorizada de los datos

4. La destrucción no autorizada de los datos.

RECUPERANDO DATOS
Restricciones

A) **Integridad de la claves**: Ningún campo de una clave candidata puede tomar **valores nulos.**

B) **Restricción de dominio:** Todo campo debe poseer valores **E** a su **dominio.**

C) **Integridad referencial:** Sea C1 un campo del archivo A1, que es clave extranjera sobre un archivo A2 (C1 y C2 = dominio). Entonces C1 debe estar instanciado siempre con **valores ya referenciados** sobre C2 o ser nulo.

RECUPERANDO DATOS
Restricciones

Problemas derivados de las restricciones

A) Inserción

Solo se puede **insertar** un registro si tiene valor de **CP no nulo.**

Solo se puede **insertar** un registro si el valor de la **CP es único** y **E al dominio.**

Solo se puede **insertar** un registro si los valores de sus **claves** extranjeras están definidos en los archivos relacionados o son nulos.

RECUPERANDO DATOS
Restricciones

Problemas derivados de las restricciones

B) Modificación

Si el campo es primo **NO** puede instanciarse con un valor **nulo**

Si el campo es primo **NO** puede instanciarse con un valor **existente** que haga que la **CLAVE CANDIDATA** se repita

Si el campo es **clave extranjera** solo puede cambiarse por un **valor nulo o por un valor existente** sobre la CP de la tabla relacionada.

RECUPERANDO DATOS
Restricciones

Problemas derivados de las restricciones

C) Borrado

Debe controlarse la integridad referencial.

Actualizar registros en cascada:

Procedimiento que cambia un valor del campo clave de la tabla principal, que automáticamente cambiará el valor de la clave extranjera de los registros relacionados en el archivo secundario.

Eliminar registros en cascada:

Procedimiento que elimina un registro de la tabla principal, automáticamente se borran también los registros relacionados en la tabla secundaria.

ORGANIZACIONES BASICAS

Archivo HEAP (montículo) o PILA: aquellos que no están ordenados

Archivos ORDENADOS: aquellos que sus registros están ordenados por algún/os campo/s particular. También se los llama SECUENCIALES.

Archivos DIRECTOS (hash): aquellos en que es posible el acceso de forma directa a un registro a partir de una clave y una f(x) de mapeo.

ORGANIZACIONES AVANZADAS

Recuperación:

Directa por Dispersión (hash)

Por Indices - (claves primarias y secundarias)

ORGANIZACIONES AVANZADAS

Acceso Directo por dispersion

Por dispersión (hash)

Cuando es necesario el acceso directo se utiliza el método de transformación de claves hashing. Se refiere al proceso de obtener una dirección de almacenamiento a partir de un campo clave. Se diseña una función para transformar un valor de la clave en otro valor que sirva como una dirección de almacenamiento.

La **división entre un primo** es uno de los posibles tipos de funciones de dispersión, donde se utiliza el resto a los efectos de mapear la clave.

ORGANIZACIONES AVANZADAS

Acceso Directo por dispersion

Requerimientos para los algoritmos de hashing

Posibilidad de repetición: La capacidad de almacenar un registro mediante un algoritmo y recuperarlo, utilizando el mismo algoritmo, es un requerimiento importante. ·

Distribución uniforme: Si se va a almacenar un archivo en un espacio que permite el almacenamiento de 10000 registros, estos deben distribuirse uniformemente en todo el espacio asignado en vez de acumularse todos juntos.

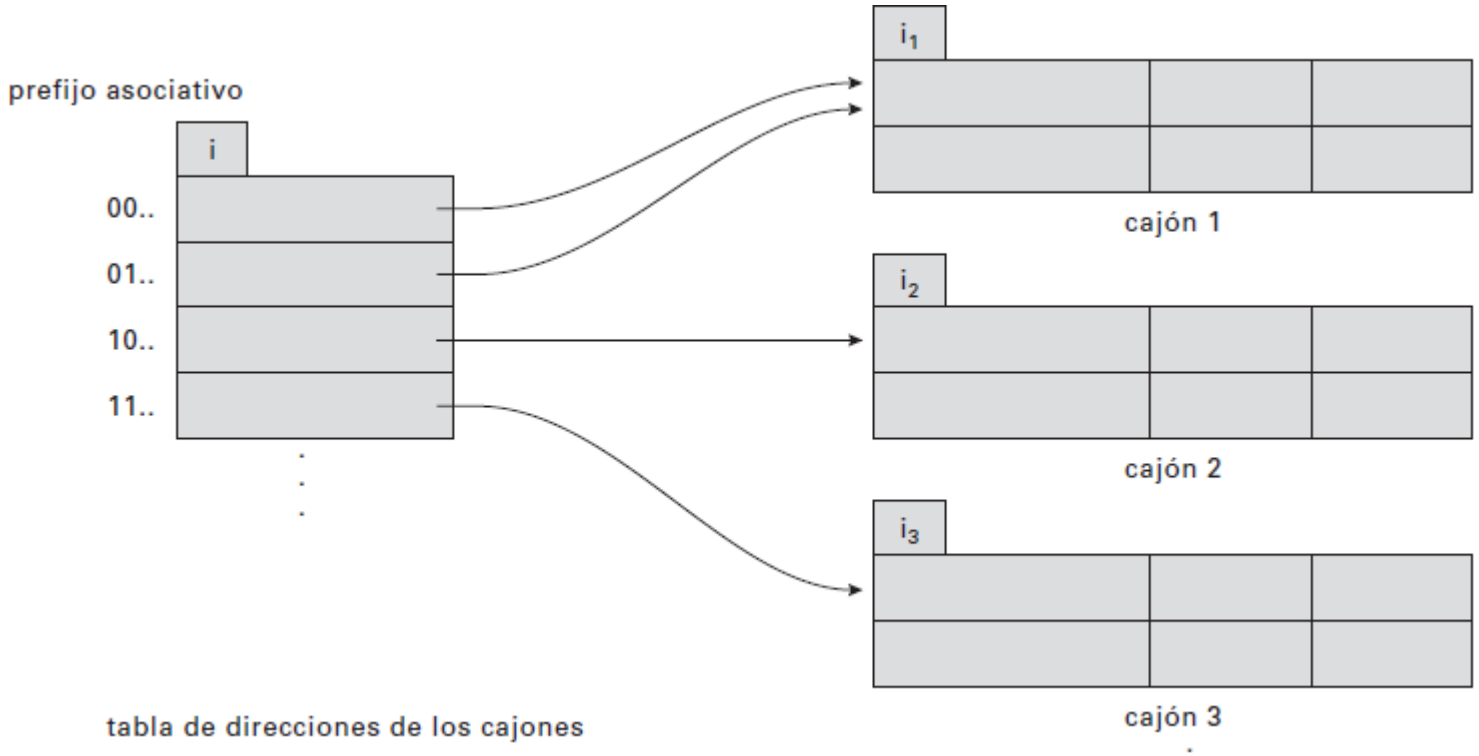
Minimizar sinónimos En la práctica, los sinónimos aparecen cuando el procedimiento de dispersión se aplica a claves distintas y produce la misma dirección de almacenamiento.

Organizaciones avanzadas
Acceso directo por dispersión
Dinámico

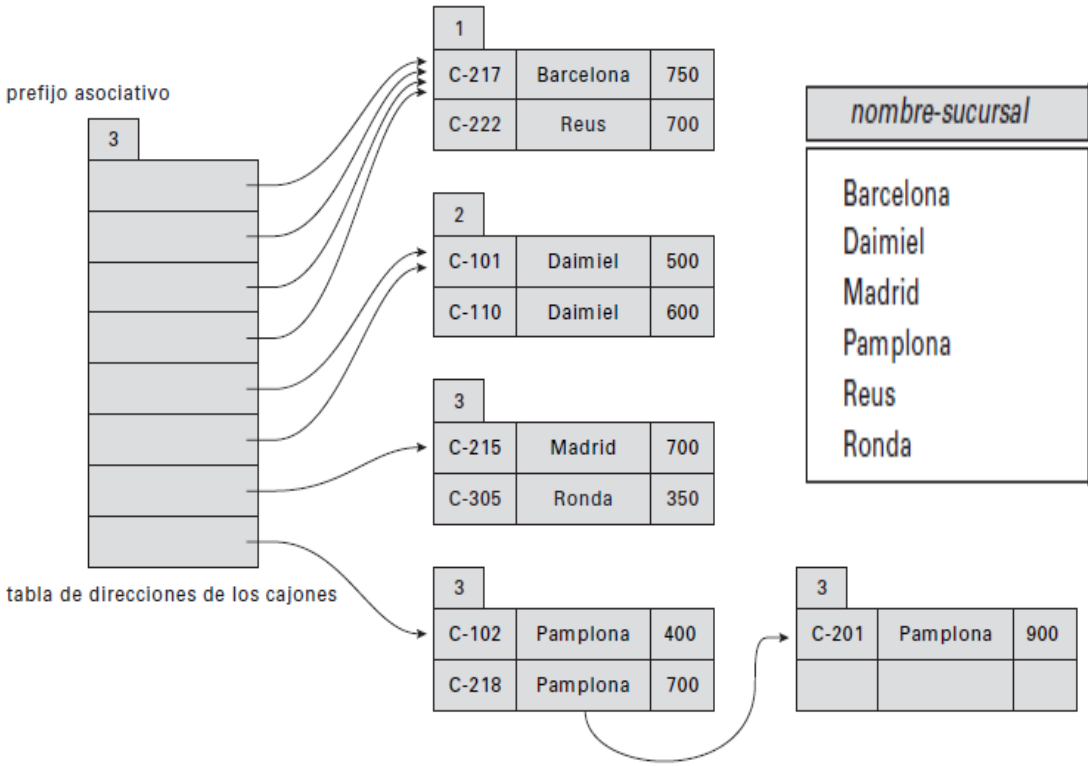
C-217	Barcelona	750
C-101	Daimiel	500
C-110	Daimiel	600
C-215	Madrid	700
C-102	Pamplona	400
C-201	Pamplona	900
C-218	Pamplona	700
C-222	Reus	700
C-305	Ronda	350

nombre-sucursal	*h(nombre-sucursal)*
Barcelona	0010 1101 1111 1011 0010 1100 0011 0000
Daimiel	1010 0011 1010 0000 1100 0110 1001 1111
Madrid	1100 0111 1110 1101 1011 1111 0011 1010
Pamplona	1111 0001 0010 0100 1001 0011 0110 1101
Reus	0011 0101 1010 0110 1100 1001 1110 1011
Ronda	1101 1000 0011 1111 1001 1100 0000 0001

Organizaciones avanzadas
Acceso directo por dispersión
Dinámico



Organizaciones avanzadas
Acceso directo por dispersión
Dinámico



nombre-sucursal	*h(nombre-sucursal)*
Barcelona	0010 1101 1111 1011 0010 1100 0011 0000
Daimiel	1010 0011 1010 0000 1100 0110 1001 1111
Madrid	1100 0111 1110 1101 1011 1111 0011 1010
Pamplona	1111 0001 0010 0100 1001 0011 0110 1101
Reus	0011 0101 1010 0110 1100 1001 1110 1011
Ronda	1101 1000 0011 1111 1001 1100 0000 0001

ORGANIZACIONES AVANZADAS

Recuperación por índices

Tipo de Índices	Descripción	Nro. De entradas del índice (Primer-Nivel)	Denso o No Denso
Primario	Campo clave de ordenación sin repeticiones	Igual al numero de bloques del archivo de datos	No Denso
Agrupación	Campo clave de ordenación con repeticiones	Igual al numero de valores diferentes del campo índice	No Denso
Secundario (Clave)	No es clave de ordenación y sin repeticiones	Igual al número de registros del archivo de datos	Denso
Secundario (No Clave)	No es clave de ordenación y con repeticiones	Igual al numero de registros o al numero de valores diferentes del campo índice	Denso

ORGANIZACIONES AVANZADAS

Recuperación por índices

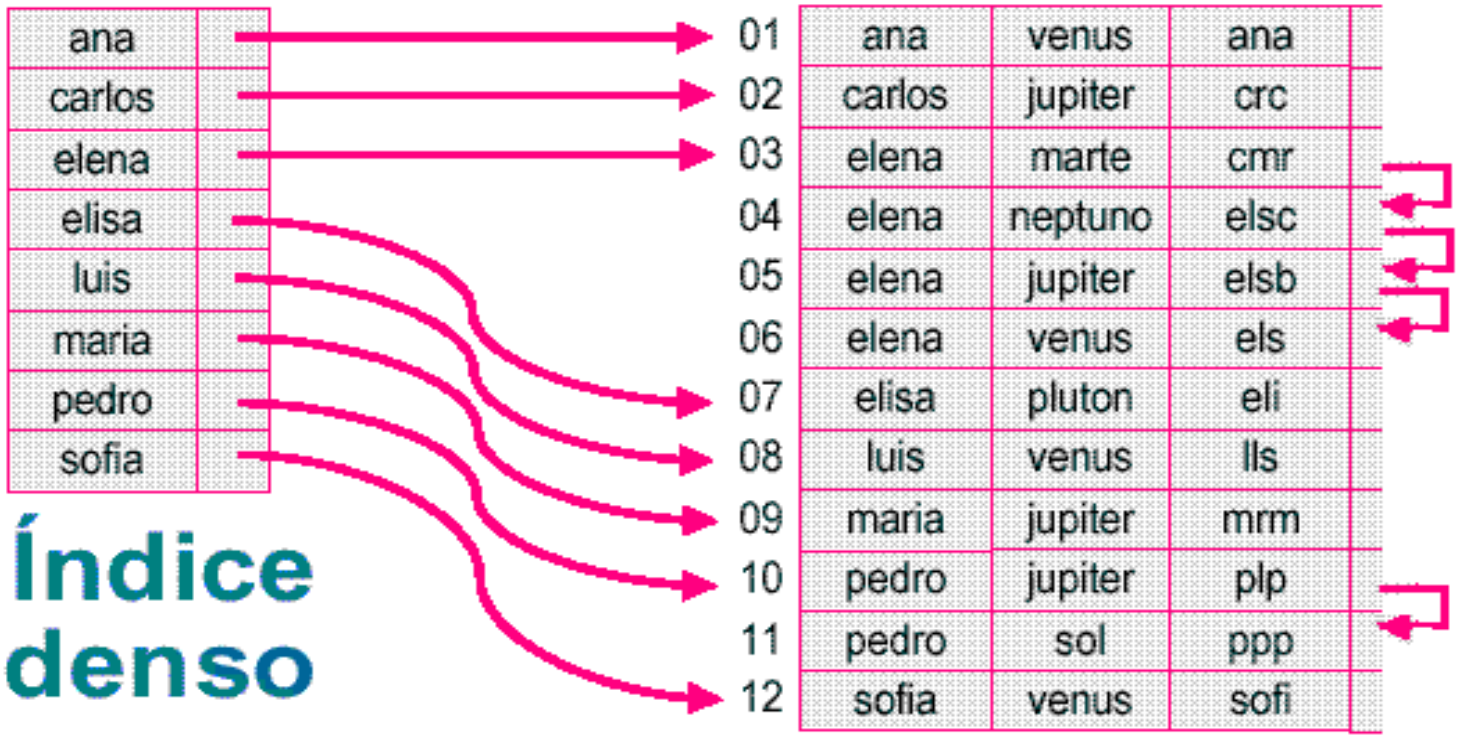
INDICE: es una estructura de datos auxiliar que ayuda en la localización de datos bajo una cierta condición de selección.

CLAVE DE BUSQUEDA: conjunto de uno o más campos del archivo para los que se construye el índice. Las entradas de datos del índice nos permiten localizar los registros de datos que tienen un valor de clave de búsqueda concreto.

ORGANIZACIONES AVANZADAS

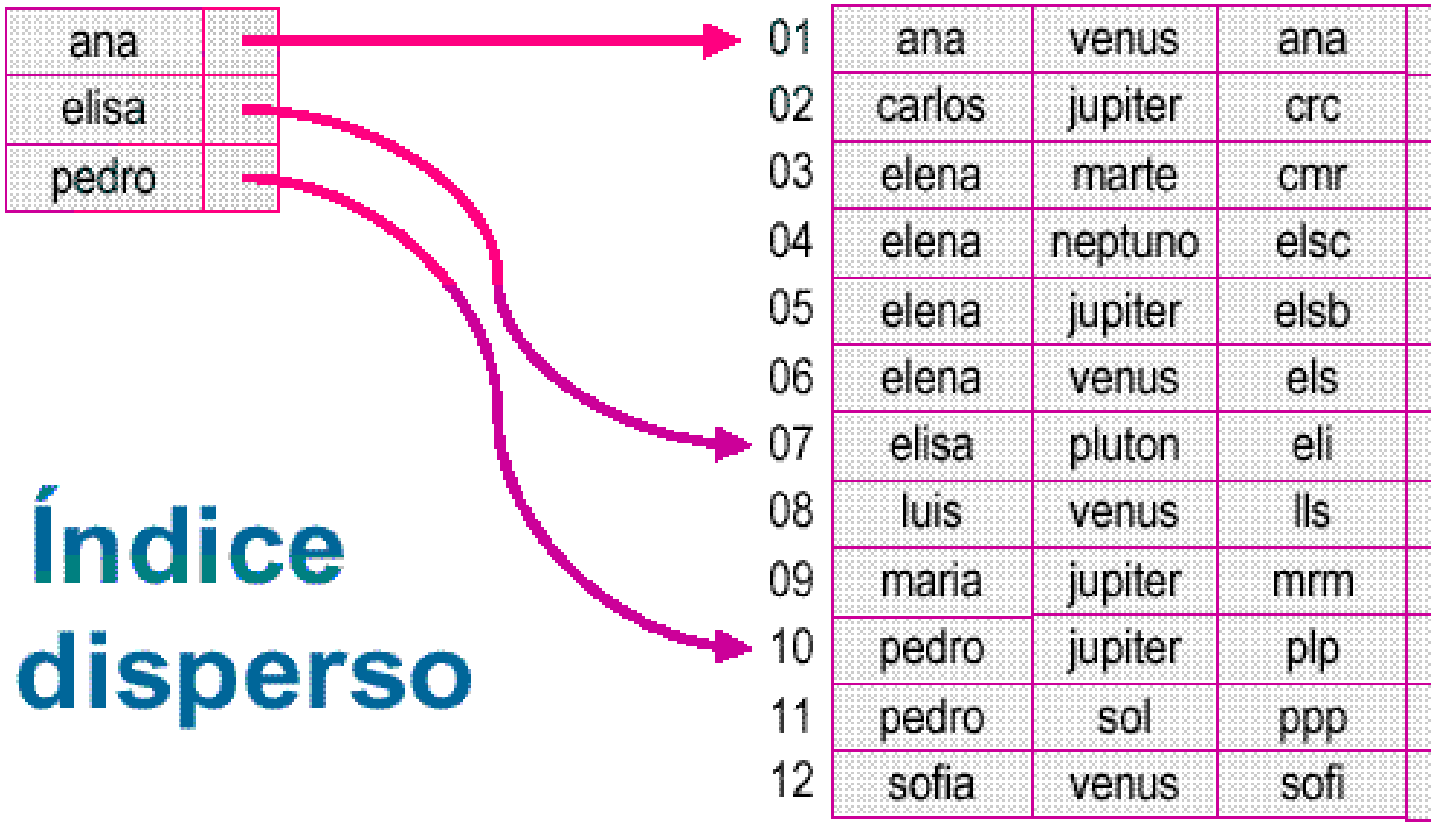
Recuperación por índices

Densos (dense) (Están todos los valores de la clave)
Dispersos (sparse) (Están algunos valores)



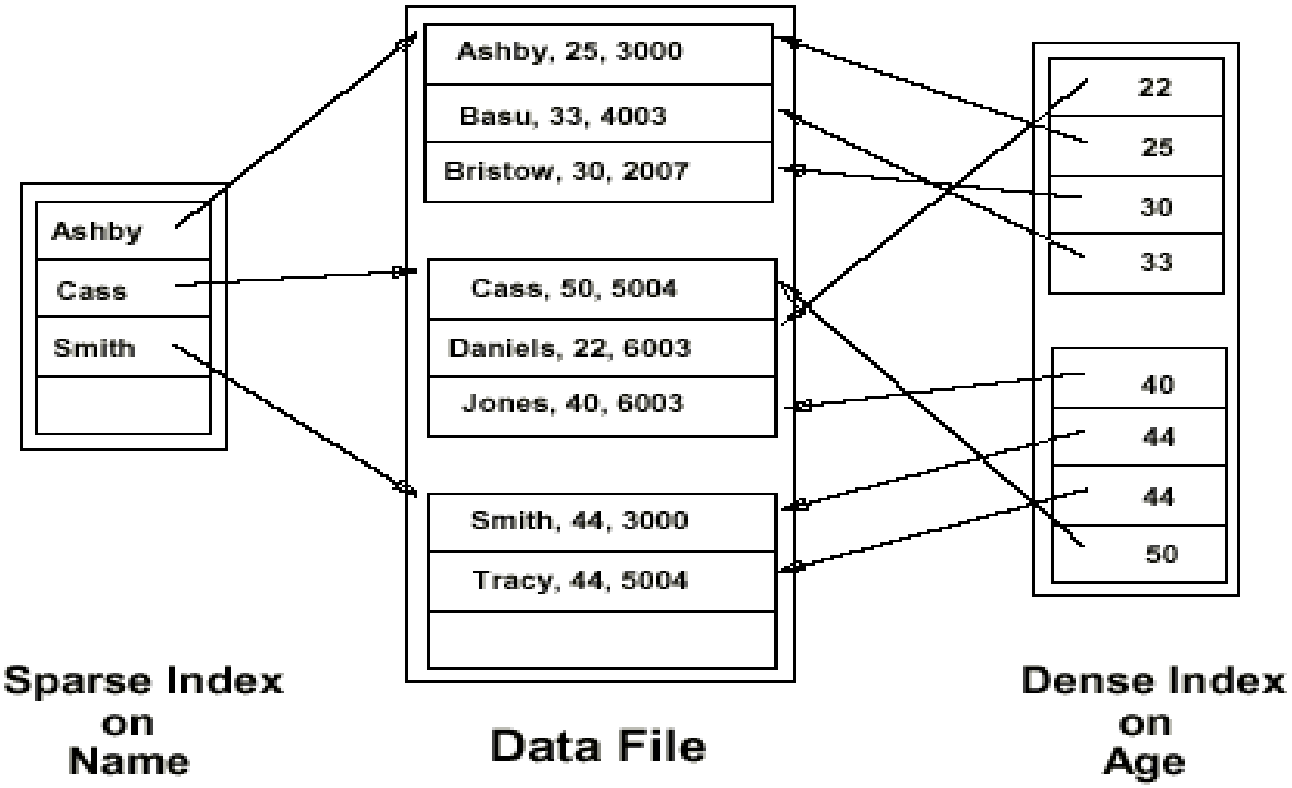
ORGANIZACIONES AVANZADAS

Recuperación por índices

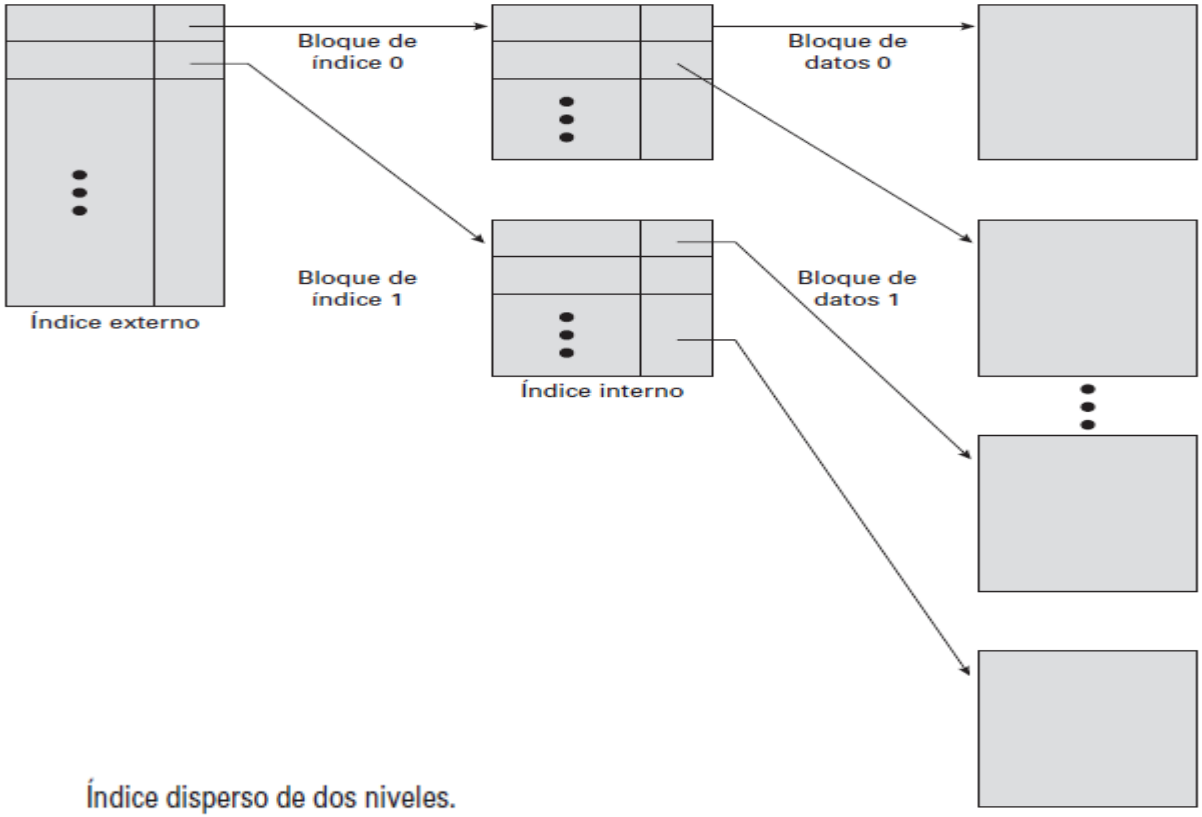


ORGANIZACIONES AVANZADAS

Recuperación por índices

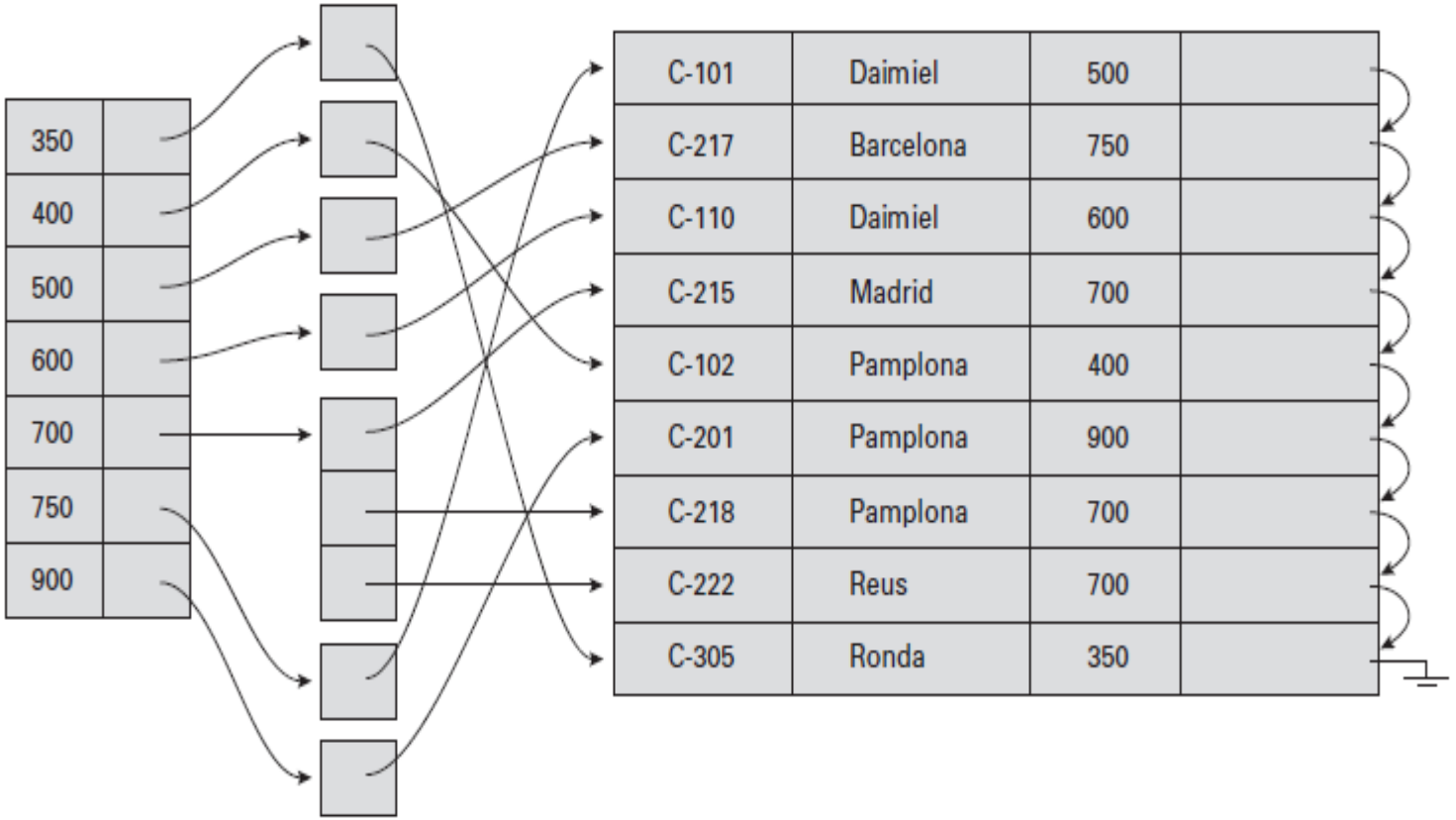


Organizaciones avanzadas



Índice disperso de dos niveles.

Organizaciones avanzadas
Índices Secundarios



Índice secundario del archivo *cuenta*, con la clave no candidata *saldo*.

Organizaciones avanzadas
Árbol B+

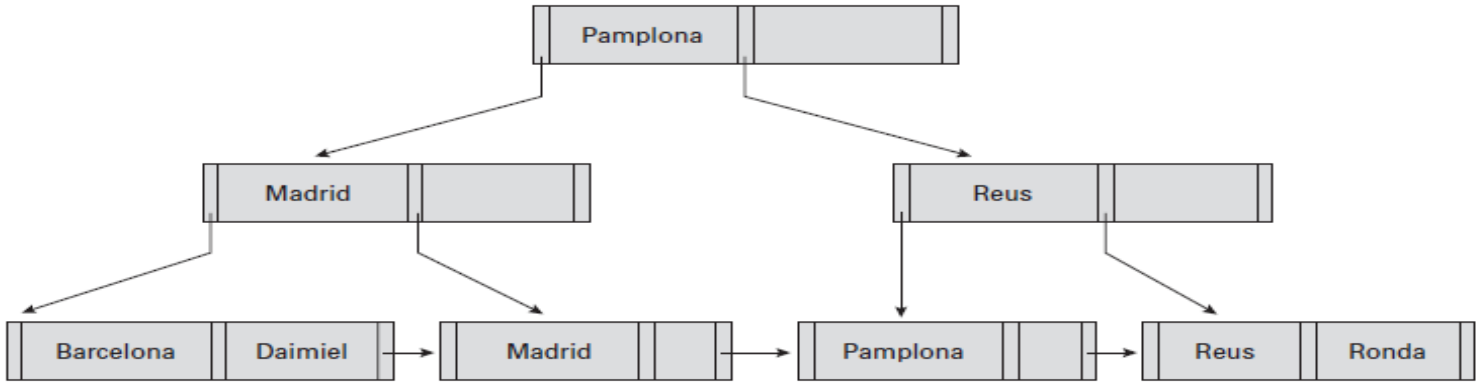
K_i son los valores de la clave de búsqueda.

P_i son los punteros a los hijos (para nodos no hoja) o a los registros o cajones de registros (para nodos hoja).

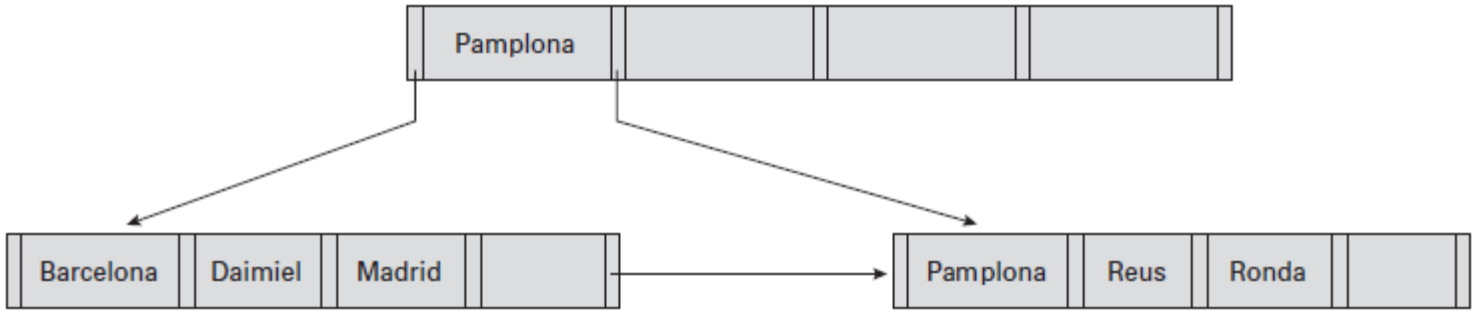
En un nodo las claves de búsqueda están ordenadas.

$$K1 < K2 < K3 < \ldots < Kn–1$$

Organizaciones avanzadas
Árbol B+



Árbol B+ del archivo *cuenta (n = 3)*



Árbol B+ del archivo *cuenta (n = 5)*