

Enunciado y solución

Ejercicio 5: Ficheros - Enunciado

La Empresa de Servicios Informáticos SII necesita optimizar los accesos a una de las tablas (EMPLEADOS) de su base de datos al ser una de las más consultadas. La tabla contiene una cardinalidad de 1.000.000 (10^6) tuplas, y cada una de ellas tiene un volumen medio de 300B. La densidad de registro es de un 94%, ya que el almacenamiento se realiza con registros de longitud variable en un soporte direccionado con $T_{bq} = 2\text{kB}$.

El esquema de relación de EMPLEADOS se representa a continuación: EMPLEADOS (DNI, nombre-c, dirección, localidad, cod_postal, categoría)

Para poder optimizar los accesos a esta tabla se ha realizado una auditoria sobre la misma.

Ejercicio 5: Ficheros - Enunciado

Existen numerosos procesos de todo tipo y naturaleza, pero los procesos críticos (a optimizar) son los siguientes:

	Atributos	Frecuencia	Descripción	Filas resultado
P_1	DNI	25%	select * from EMPLEADOS where DNI=x;	1
P_2	localidad, cod_postal	50%	select * from EMPLEADOS where localidad=x and cod_postal= y;	50
P_3	cod_postal	25%	Update empleados set categoria= z where cod_postal=x;	100

La organización de la que partimos es O_0 : organización serial no consecutiva, con $E_c=4$ y espacio libre distribuido de 10% (espacio suficiente para realizar las modificaciones). Las organizaciones que se plantean como posibles mejoras son:

- O_1 : Direccionada sobre $CD=DNI$, con función de transformación sobre $N=5 \cdot 10^4$ que produce una tasa de 0.01% de registros desbordados, gestionados en área de desbordamiento serial.
- O_2 : Secuencial (mismo cubo; inserción en área desordenada; CO a elegir por el alumno).

Ejercicio 5: Ficheros - Enunciado

- a) Compare el coste global (en accesos) de las organizaciones candidatas (O_0 , O_1 , y O_2) teniendo en cuenta que se debe elegir la clave de ordenación de O_2 . Calcule las densidades de cada organización (ideal, real y de ocupación) y justifique las decisiones tomadas, comentando las ventajas e inconvenientes de cada una de las organizaciones.
- b) A cada organización se le puede añadir algún índice para mejorar el rendimiento. Tómese como tamaño de punteros interno y externo 4 B (el externo contiene partes alta y baja). Elija para cada organización el índice (o índices) denso(s) que estima más adecuado(s) y justifique porqué. Calcule los costes de estas nuevas organizaciones y explique cuál es la mejor.

** Las longitudes de los atributos son: DNI tiene 9 B de tamaño fijo; cod_postal tiene 5 B de tamaño fijo; localidad, tamaño variable, de media 55B y cardinalidad de 800 valores distintos; categoría es de tamaño variable, de media 3B, con 9 registros por valor.

Ejercicio 5: Ficheros – (a) – O₀

O₀ serial no consecutiva, habrá que calcular:

- El tamaño de cubo
- El tamaño del fichero en cubos
- Las densidades ideal, real y de ocupación
- Costes

$$T_c = \frac{T_{bq} * E_c * (1 - ELD)}{\text{Volumen Real}} = \frac{2048 \text{ bytes/bq} * 4 \text{ bq/cubo} * (1 - 0,1)}{300 \text{ bytes/registro}} = 24,57 \rightarrow \mathbf{24 \text{ reg/cubo}}$$

$$T_{\text{fichero}} = \frac{\text{número de registros}}{T_c} = \frac{10^6 \text{ registros}}{24 \frac{\text{registros}}{\text{cubo}}} = \mathbf{41667 \text{ cubos}}$$

$$d_i = 94\%$$

$$d_r = \frac{\text{numReg} * \text{Inf útil}}{T_{\text{fichero}} * E_c * T_{bq}} = \frac{10^6 \text{ registros} * 300 \text{ bytes/reg} * 0,94}{41667 \text{ cubos} * 4 * 2048 \text{ bytes/bq}} = 0,8262 \rightarrow \mathbf{82,62\%}$$

$$d_o = \frac{\text{núm. regist.}}{N * T_c} = \frac{10^6 \text{ registros}}{41667 \text{ cubos} * 24 \text{ reg/cubo}} = 0,9999 \rightarrow \mathbf{99,99\%}$$

Ejercicio 5: Ficheros – (a) – O₀

P₁: Consulta por DNI con resultado de 1, por tanto identificativa

Leer hasta encontrar el cubo donde está el registro: Caso mejor/peor

$$C(O_0, P_1) = \left\lceil \frac{1 + 41667}{2} \right\rceil * E_c = 83336 \text{ accessos}$$

P₂: Consulta por localidad y código postal no identificativa (50 registros)

Leer de principio a fin – Full Scan

$$C(O_0, P_2) = [\text{Número de cubos}] * E_c = 166668 \text{ accessos}$$

P₃: Actualización por código postal, se actualiza categoría (100 registros)

Leer de principio a fin – Full Scan, el caso peor es encontrar cada registro en un cubo diferente, por lo que habrá que reescribir 100 de esos cubos.

$$C(O_0, P_3) = \text{Sel} + \text{act} = (41667 + 100) * E_c = 167068 \text{ accessos}$$

$$C(O_0) = f_1 * C(O_0, P_1) + f_2 * C(O_0, P_2) + f_3 * C(O_0, P_3) = 145935 \text{ accesos}$$

Ejercicio 5: Ficheros – (a) – O_1

O_1 direccionada por CD=DNI sobre $N=5*10^4$, gestión de desbord. en área indpte de organización serial. Tasa de desbordamientos del 0,01% del total de registros

- El tamaño de cubo
- Tamaño del área independiente serial
- El tamaño del fichero en cubos
- Las densidades ideal, real y de ocupación
- Costes

$T_c \rightarrow$ Igual que en O_0

Además de los 50000 cubos necesarios para el direccionamiento y a los que irán a parar un 99,99% de los registros, tiene que haber un área independiente para el 0,01% de registros que encuentran cubos llenos (0,01% de 10^6 , 100 registros)

$$T_{\text{area desb}} = \frac{\text{nº de registros desb.}}{T_c} = \frac{100 \text{ registros}}{24 \frac{\text{registros}}{\text{cubo}}} = \mathbf{5 \text{ cubos}} \text{ (redondear arriba)}$$

$$T_{\text{fichero}} = \mathbf{50000 + 5 \text{ cubos}}$$

Ejercicio 5: Ficheros – (a) – O₁

$$d_i = 94\%$$

$$d_r = \frac{\text{numReg} * \text{Inf útil}}{T_{\text{fichero}} * E_c * T_{\text{bq}}} = \frac{10^6 \text{ registros} * 300 \text{ bytes/reg} * 0,94}{5 * 10^4 + 5 \text{ cubos} * 4 * 2048 \text{ bytes/bq}} = 0,6884 \rightarrow \mathbf{68,84\%}$$

$$d_o = \frac{\text{núm. regist.}}{N * T_c} = \frac{10^6 \text{ registros}}{5 * 10^4 + 5 \text{ cubos} * 24 \text{ reg/cubo}} = 0,8332 \rightarrow \mathbf{83,32\%}$$

P₁: Consulta por DNI con resultado de 1, por tanto identificativa

El 99,99% de las veces el direccionamiento permitirá encontrarlo en un acceso, pero 1 de cada 10000 veces además de ese acceso habrá que ir al área de desbordamiento y buscar, teniendo en cuenta que es identificativa (caso mejor/peor)

$$C(O_1, P_1) = \left[0,9999 * 1 + 0,0001 * \left(1 + \frac{1 + 5}{2} \right) \right] * E_c = 4,0012 \text{ accessos}$$

Ejercicio 5: Ficheros – (a) – O₁

P₂: Consulta por localidad y código postal no identificativa (50 registros)

Leer de principio a fin – Full Scan, incluida la zona desordenada

$$C(O_1, P_2) = [5 * 10^4 + 5] * E_c = 200020 \text{ accessos}$$

P₃: Actualización por código postal, se actualiza categoría (100 registros)

Leer de principio a fin – Full Scan (incluida el área de desbordamiento), el caso peor es encontrar cada registro en un cubo diferente, por lo que habrá que reescribir 100 de esos cubos.

$$C(O_1, P_3) = Sel + act = (5 * 10^4 + 5 + 100) * E_c = 200420 \text{ accessos}$$

$$C(O_1) = f_1 * C(O_1, P_1) + f_2 * C(O_1, P_2) + f_3 * C(O_1, P_3) = 150116 \text{ accesos}$$

Ejercicio 5: Ficheros – (a) – O_2

O_2 secuencial con clave de ordenación a elegir por el alumno

- Elección de la clave de ordenación
- El tamaño de cubo
- El tamaño del fichero en cubos
- Tamaño del área independiente serial
- Las densidades ideal, real y de ocupación
- Costes

Elección del código postal como clave de ordenación, pues nos valdría para dos consultas: para el proceso 3 de manera directa, para el proceso 2 requeriría de filtrado posterior para la segunda condición (localidad)

$T_c \rightarrow$ Igual que en O_0

$T_{\text{fichero}} \rightarrow$ Igual que en O_0

$d_r \rightarrow$ Igual que en O_0

$d_o \rightarrow$ Igual que en O_0

Ejercicio 4: Ficheros – (a) – O₂

P₁: Consulta por DNI con resultado de 1, por tanto identificativa

Leer hasta encontrar el cubo donde está el registro: Caso mejor/peor

$$C(O_2, P_1) \rightarrow \text{Igual que en } O_0$$

P₂: Consulta por localidad y código postal no identificativa (50 registros)


Se puede aplicar la búsqueda dicotómica. Hay que tener en cuenta que cuando se busca por código postal sobre 10^6 registros y cada valor de código postal devuelve en término medio 100 registros, la búsqueda se hace sobre 10000 valores. Una vez encontrada hay que desplazarse para encontrar los 100 registros con el código postal. Una vez recuperados los 100 registros, se puede filtrar para quedarnos con los 50 que además cumplen con la localidad

$$C(O_2, P_2) = \left[\log_2(10^4 + 1) + \left(\frac{100 + 1}{24} \right) \right] * E_c = 76 \text{ accesos}$$

Ejercicio 4: Ficheros – (a) – O₂

P₃: Actualización por código postal, se actualiza categoría (100 registros)

Se puede aplicar la búsqueda dicotómica para la búsqueda. Hay que tener en cuenta que cuando se busca por código postal sobre 10⁶ registros y cada valor de código postal devuelve en término medio 100 registros, la búsqueda se hace sobre 10000 valores. Una vez encontrada hay que desplazarse para encontrar los 100 registros con el código postal. La actualización no supone alterar el orden, pues se modifica la categoría, no el código postal, habría que reescribir los cubos con los registros modificados

$$C(O_2, P_3) = \left[\log_2(10^4 + 1) + \left(\frac{100 + 1}{24} \right) + \frac{k}{T_c} \right] * E_c = (19 + 5) * E_c = \mathbf{96 \text{ acc.}}$$


$$C(O_1) = f_1 * C(O_2, P_1) + f_2 * C(O_2, P_2) + f_3 * C(O_2, P_3) = \mathbf{20896 \text{ accesos}}$$

Ejercicio 5: Ficheros – Enunciado (b)

- a) A cada organización se le puede añadir algún índice para mejorar el rendimiento. Tómese como tamaño de punteros interno y externo 4 B (el externo contiene partes alta y baja). Elija para cada organización el índice (o índices) denso(s) que estima más adecuado(s) y justifique porqué. Calcule los costes de estas nuevas organizaciones y explique cuál es la mejor.

** Las longitudes de los atributos son: DNI tiene 9 B de tamaño fijo; cod_postal tiene 5 B de tamaño fijo; localidad, tamaño variable, de media 55B y cardinalidad de 800 valores distintos; categoría es de tamaño variable, de media 3B, con 9 registros por valor.

Ejercicio 5: Ficheros – Enunciado (b)

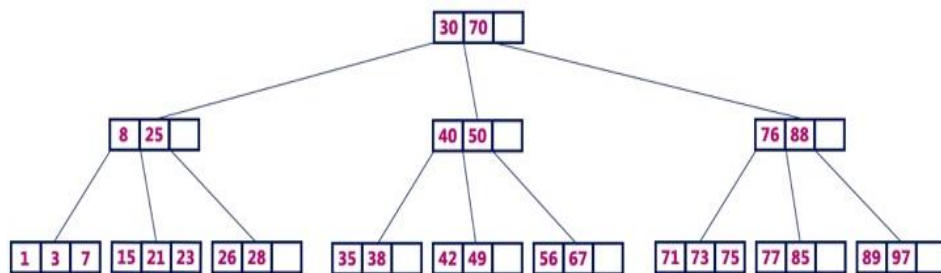
- a) A cada organización se le puede añadir algún índice para mejorar el rendimiento. Tómese como tamaño de punteros interno y externo 4 B (el externo contiene partes alta y baja). Elija para cada organización el índice (o índices) denso(s) que estima más adecuado(s) y justifique porqué. Calcule los costes de estas nuevas organizaciones y explique cuál es la mejor.

** Las longitudes de los atributos son: DNI tiene 9 B de tamaño fijo; cod_postal tiene 5 B de tamaño fijo; localidad, tamaño variable, de media 55B y cardinalidad de 800 valores distintos; categoría es de tamaño variable, de media 3B, con 9 registros por valor.

- Se valora un índice B con CI = DNI (identificativo)
- Se valora un índice B+ con CI = CP (no identificativo)

Ejercicio 5: Ficheros – (b) – B sobre DNI

¿Cómo son los nodos y el árbol?



$$m \cdot T_{\text{ptro_interno}} + k \cdot T_{\text{entrada}} < T_{\text{nodo}}$$

$$m \cdot 4 + (m - 1) \cdot (4 + 9) < 2048$$

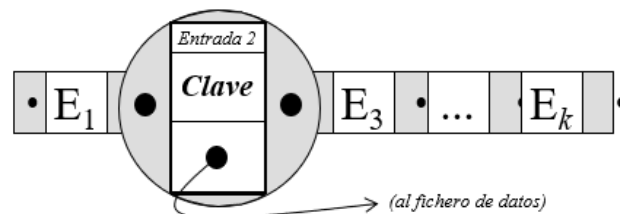
$$m \leq 2061/17$$

$$m = 121$$

$$k = 120 \text{ (puesto que } k = m - 1)$$

$$K_{\text{mín}} = k/2 = 60$$

$$m_{\text{mín}} = (m+1)/2 = 61$$



Cada nodo:

- Contiene entradas de índice (pares clave indización-puntero a los datos)
- Punteros (para apuntar nodos hijo)

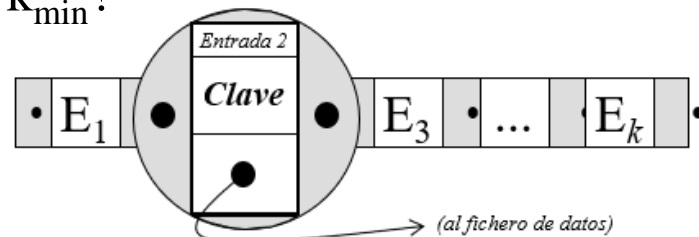
Consideraciones

- Puntero interno de 4B
- Puntero externo de 4B
- Tamaño de la entrada: 4B del puntero externo + 9B fijos de la clave (DNI)
- $k = m - 1$
- Tomamos el tamaño de nodo más pequeño posible, es decir 1 bloque (2048 B)

Ejercicio 5: Ficheros – (b) – B sobre DNI

¿Cuántos bloques necesitamos considerando la k_{\min} ?

$$T_f(\text{índ}_K) = 10^6/60 = 16667 \text{ bloques}$$



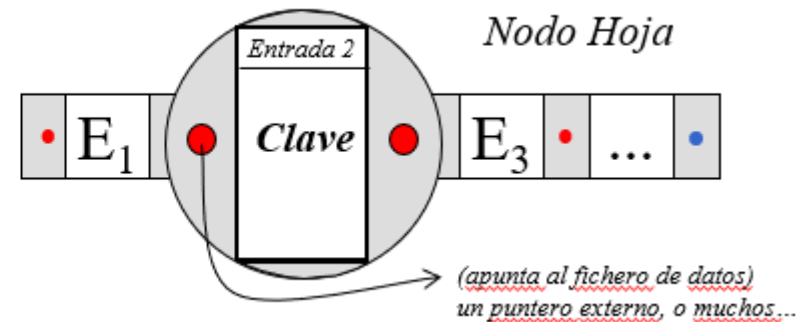
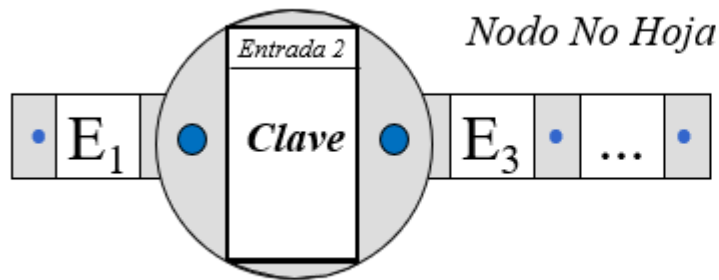
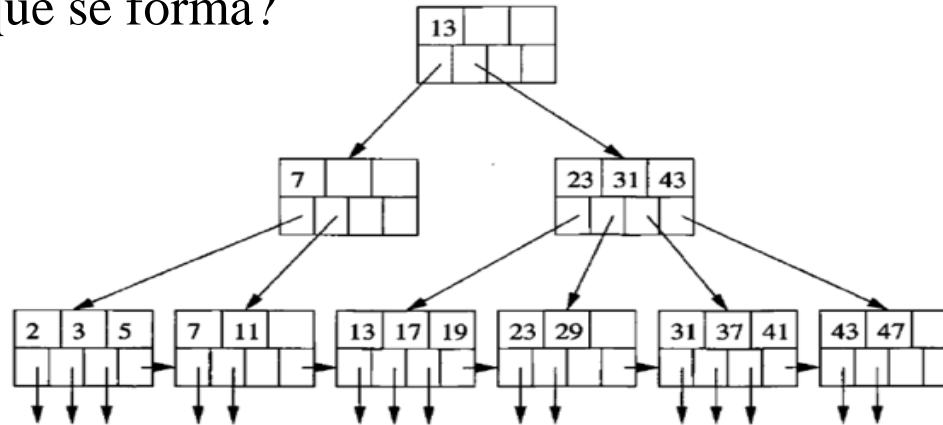
¿Cuántos niveles necesitamos en el árbol?

Nivel	nodos	entradas	Acumulado	
1	1	1	1	
2	2	$2*60=120$	121	
3	$2*61=122$	$122*60=7320$	7441	
4	$122*61=7442$	$7442*60=446520$	453961	
5	$7442*61=453962$	$453962*60=27237720$	27691681	$> 10^6$

$n = 4$ niveles, en general 3 accesos si la raíz está en memoria

Ejercicio 5: Ficheros – (b) – B⁺ sobre CP

¿Cómo es el árbol que se forma?



- En los nodos con **hijos** (**no hoja**) no hay punteros externos
- En los nodos **hoja** no hay punteros a nodo hijo, pero sí habrá punteros externos. Para apuntar a los datos, la entrada debe estar en una hoja, por lo que todas las entradas están en nodos hojas, y en los no hoja sólo hay copias discriminantes

Ejercicio 5: Ficheros – (b) – B⁺ sobre CP

Orden del árbol (m): se calcula para nodos no hoja, como en árboles B, teniendo en cuenta que las entradas de esos nodos carecen de puntero externo (CP ->5B)

$$m \cdot T_{\text{puntero_interno}} + (m-1) \cdot T_{\text{clave}} < T_{\text{nodo}}$$

$$m \cdot 4 + (m-1) \cdot 5 < 2048 \rightarrow m = 228 \rightarrow m_{\min} = (m+1)/2 = 114$$

Tamaño de las entradas (marca por el n° de reg. con ese CP):

$$T_{\text{entrada}} = \text{Tamaño CP} + \text{Marca} + \text{Tamaño puntero} \cdot n^{\circ} \text{registros}$$

$$T_{\text{entrada}} = 5B + 1B + 4B \cdot 100 = 406 B$$

Cálculo del valor de k:

$$k \cdot T_{\text{entrada}} + T_{\text{puntero_interno}} < T_{\text{nodo}}$$

$$k \cdot 406 + 4 < 2048 \rightarrow k = 5 \rightarrow k_{\min} = (k+1)/2 = 3$$

¿Cuántas hojas tendremos?

¿Hay que considerar que tenemos 10000 valores distintos de CP porque se recuperan unos 100 por valor

Número de hojas = $10000 / 3 = 3333$ hojas (que constituyen el nivel n)

¿Cuántos nodos necesitamos en el nivel $n-1$?

Tendremos que considerar el m_{\min} , puesto que los nodos de este nivel serán no hoja $\rightarrow 3333/114=29$

¿Cuántos nodos necesitamos en el nivel $n-2$?

Tendremos que considerar el m_{\min} , puesto que los nodos de este nivel serán no hoja $\rightarrow 29/114 < 1 \rightarrow$ éste será el raíz

Por tanto, el árbol tendrá 3 niveles (2 accesos a memoria con la raíz siempre)

Ejercicio 5: Ficheros – (b) – Impacto O_0

El B ayuda a la consulta identificativa y el B+ con las consultas no identificativa

	P_1 Selección por clave identificativa DNI	P_2 Selección por clave no identificativa (50 registros)	P_3 Actualización por clave no identificativa CP (100 registros)	Coste
O_0	83336	166668	176078	145935
O_0'	Usar B Lectura B: $(n-1) = 3$ Leer cubo $1 * E_c = 4$ 7	Usar B+ $(n-1) = 2$ (no se ve modificado) Leer cubo suponiendo cada uno de diferente (después filtrar para Localidad) $100 * E_c = 400$ 402	Usar B+ $(n-1) = 2$ (no se ve modificado) Leer cubos (supuestos en dist. cubos) y escribirlos $100 * E_c = 400$ $100 * E_c = 400$ 802	403,25
f_i	25% → 0,25	50% → 0,50	25% → 0,25	

Ejercicio 5: Ficheros – (b) – Impacto O_1

El B+ con las consultas no identificativa

	P_1 Selección por clave identificativa DNI	P_2 Selección por clave no identificativa (50 registros)	P_3 Actualización por clave no identificativa CP (100 registros)	Coste
O_1	4,0012	200020	200420	150116
O_1'	<p>No B porque no lo mejoraría</p> <p>Se mantiene el coste</p> <p>4,0012</p>	<p>Usar B+ (n-1) = 2 (no se ve modificado)</p> <p>Leer cubo suponiendo cada uno de diferente (después filtrar para Localidad) 100*Ec = 400</p> <p>402</p>	<p>Usar B+ (n-1) = 2 (no se ve modificado)</p> <p>Leer cubos (supuestos en dist. cubos) y escribirlos 100*Ec = 400 100*Ec = 400</p> <p>802</p>	402,5003
f_i	25% → 0,25	50% → 0,50	25% → 0,25	

Ejercicio 5: Ficheros – (b) – Impacto O_2

El B ayuda a la consulta identificativa

	P_1 Selección por clave identificativa DNI	P_2 Selección por clave no identificativa (50 registros)	P_3 Actualización por clave no identificativa CP (100 registros)	Coste
O_2	83336	76	96	20896
O_2'	Usar B Lectura B: $(n-1) = 3$ Leer cubo $1 * E_c = 4$ 7	No B+ Porque no lo mejoraría 76	No B+ Porque no lo mejoraría 96	63,75
f_i	25% → 0,25	50% → 0,50	25% → 0,25	