Titulación: GRADO INGENIERIA INFORMATICA Año Académico: 2019/2020 -- Curso: 2º Asignatura: Ficheros y bases de datos Ejercicio F5 – Ejercicio de Estructuras de Ficheros	uc3m | Universidad Carlos III de Madrid

1. - 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} = 2KB.

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

Proceso	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 $N=5*10^4$ 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).

2. – APARTADOS A REALIZAR

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.

a.) $T_c = \frac{(4bq \cdot 2kB/b - 0) \cdot 0'9}{T_b = 300B}$ (ELD) $= 24$ reg/cubo. $\quad N = \frac{10^6 \text{ regs}}{24} = 41667$ cubos.

Consulta identificativa

$C(O_0, P_1) = \frac{N+1}{2} = 20834$ acc. cubo $\cdot 4 = 83336$ accesos

Consulta no identificativa

$C(O_0, P_2) = N = 41667$ acc. cubo $\cdot 4 = 166.668$ accesos a bloque.

$C(O_0, P_3) = N + \text{regs. mod.} = 41667 + 100$ acc. cubo $\cdot 4 = 167.068$ acc

$C(O_0, P) = 0'25 \cdot 83336 + 0'5 \cdot 166668 + 0'25 \cdot 167.068 = 145935$

CDisp. = DNI $\quad N = 50000 \quad T_{desb} = 0'01\,\% \quad N' = \frac{10^6 \cdot 0'0001}{24} = 5$ cubos

$P_{desb} = 0'0001$

$C(O_1, P_1) = 1 + T_{desb} \cdot \frac{N'+1}{2} = 1 + 0'0001 \cdot \frac{5+1}{2} = 1'0003$ acc. cubo $= 4$ acc.

$C(O_1, P_2) = N + N' = 50000 + 5 = 50.005$ acc. cubo $= 200\,020$ acc

$C(O_1, P_3) = N + N' + 100 = 50\,105$ acc cubo $= 200\,420$

si cambiais DNI cerca interbloq.

$C(O_1, P) = 0'25 \cdot 4 + 0'5 \cdot 200020 + 0'25 \cdot 200420 = 150\,116$

cp + sec

Secuencial $\quad CO =$ A elegir. Se escoge (cod-postal), al estar en 2 cartesas. Se añade + (tambien localidad) (solo cp tambien vale)

$N = 41667$ cubos

$C(O_2, P_1) = \frac{N+1}{2} = \frac{41667+1}{2} = 20\,834$ acc. cubo $= 83\,336$ acc. blq

min(41667, $\frac{10^6}{50}$)

$C(O_2, P_2) = \lceil \log_2(20.000+1) \rceil + \left\lceil \frac{50+1}{24} \right\rceil = 15+3 = 18$ acc. cb $= 72$ acc. blq

$C(O_2, P_2) = \lceil \log_2(10.000+1) \rceil + \left\lceil \frac{100+1}{24} \right\rceil + \frac{100}{24} = 14+5+5 = 24$ acc. cb $= 96$ acc. blq.

min (10.000, 41667)

Aprovecha secuencialidad

$C(O_2, P) = 0'25 \cdot 83336 + 0'5 \cdot 72 + 0'25 \cdot 96 = 20\,894$ acc blq

La mejor CO = cp + loc

porque admite con CO = cp

Secuencial CO = loc + cp

$C(O_2, P_1) = \frac{41667+1}{2} = 83\,336$ acc. blq.

$C(O_2, P_2) = \lceil \log_2(20001) \rceil + \lceil \frac{51}{24} \rceil = 18 \underset{cb}{acc} = 72$ acc blq.

$C(O_2, P_3) = N + 100 = 41667 + 100 = 167\,068$ acc blq.

$C(O_2, P) = 0'25 \cdot 83\,336 + 0'5 \cdot 72 + 0'25 \cdot 167\,068 = 62\,637$

$d_r(O_0) = \frac{(300 \cdot 0'94) \cdot 10^6}{41667 \cdot 4 \cdot 2048} = 82'6\,\%$ $\quad d_i = 0'94 = \frac{util}{vel = 300}$

$d_0(O_0) = 100\,\%$ Es serial, no hay huecos.

$d_r(O_1) = \frac{(300 \cdot 0'94) \cdot 10^6}{(50000+5) \cdot 4 \cdot 2048} = 68'8\,\%$

$d_0(O_1) = \frac{r - r'}{N \cdot T_c} = \frac{10^6 - 100}{50000 \cdot 24} = 83'3\,\%$

$d_0(O_2) = 100\,\%$ Es secuencial, es serial con orden, sin huecos.

b.) Arbol B sobre DNI, al ser clave identificativa.

$m \cdot 4 + k(9+4) \leq 2048$; $m = k+1$

$4k + 4 + 13k \leq 2048$; $k = 120$ $m = 121$

$k_{min} = \frac{120}{2} = 60$ $m_{min} = 61$

1.000.000 de entradas.

nivel	nodos	entradas	entradas acu.
1	1	1	1
2	2	$2 \cdot 60 = 120$	$120+1 = 121$
3	$2 \cdot 61 = 122$	$122 \cdot 60 = 7320$	7441
4	$122 \cdot 61 = 7442$	$7442 \cdot 60 = 446520$	453961 ← Tiene 4 niveles
5	$7442 \cdot 61 = 453962$	$453962 \cdot 60 = 27.237.720$	$27.691.681 > 10^6$ Demasiadas

Hay 4 niveles $\Rightarrow$ $n_1 = 4$

Arbol B^+ cp + loc

$$m \cdot 4 + (m-1)(\underbrace{5}_{cp} + \underbrace{1}_{marca\ loc} + \underbrace{55}_{loc}) \leq 2048$$

$$m = 32 \qquad m_{min} = \frac{32+1}{2} = 16$$

$$k(\underbrace{5}_{cp} + \underbrace{1}_{marca\ loc} + \underbrace{55}_{loc} + \underbrace{4 \cdot 50}_{50\ punteros\ provider}) + 4 \leq 2048$$

$$261\,k \leq 2044 \;;\; k = 7 \qquad k_{min} = \frac{7+1}{2} = 4$$

Entradas (comb. de cp y loc) $= \frac{10^6}{50} = 20\,000$ posibles valores contempladas

#nodos hoja $= \frac{20\,000}{4} = 5\,000$ hojas

$$\#\text{nodos}(n-1) = \frac{5000}{16} = 312$$

$$\#\text{nodos}(n-2) = \frac{312}{16} = 19$$

$$\#\text{nodos}(n-3) = \frac{19}{16} = 1 \quad \text{Raiz} \Rightarrow \text{Hay } n-3=1;\ n=4 \text{ niveles}$$

$$\boxed{n_2 = 4}$$

El update solo cambia categoria y no interviene en los indices.

Serial ⇒ Usamos los dos índices. < B DNI / B+ cp+loc

$C(O_0', P_1) = \overbrace{(n_1 - 1)}^{\text{Trabajo en bloque}} + \underbrace{1}_{\text{El árbol que apunta}} \cdot 4 = 4 - 1 + 4 = 7$ acc

$C(O_0', P_2) = (n_2 - 1) + 50 \cdot 4 = 4 - 1 + 200 = 203$ acc

$C(O_0', P_3)$: range scan que lee $\frac{100}{50} = 2$ entradas + 1 fallo : $(n_2 - 1) + \frac{3-1}{K\,min}$

$C(O_0', P_3) = (n-1) + \underbrace{1}_{\text{Por si está en siguiente.}} + 100 \cdot 4 + 100 \cdot 4 = 804$ acc

$C(O_0', P) = 0'25 \cdot 7 + 0'5 \cdot 203 + 804 \cdot 0'25 = 304'25$ acc

Dispersa ⇒ B+ sobre cp+loc

$C(O_1', P_1) = 1 \cdot 4 = 4$ acc

$C(O_1', P_2) = (n_2 - 1) + 50 \cdot 4 = 203$ acc

$C(O_1', P_3) = C(O_0', P_3) = 804$ acc

$C(O_1', P) = 0'25 \cdot 4 + 0'5 \cdot 203 + 804 \cdot 0'25 = 303'75$ acc

Sec. co = cp+loc ⇒ B DNI

$C(O_2', P_1) = (n_1 + 1) + 1 \cdot 4 = 7$ acc

$C(O_2', P_2) = C(O_2, P_2) = 72$

$C(O_2', P_3) = C(O_2, P_3) = 96$

$C(O_2', P) = 0'25 \cdot 7 + 0'5 \cdot 72 + 0'25 \cdot 96 = 61'75$ acc

La mejor, pero degenera mucho.