Relación 2

1. Sean las relaciones R y S con los siguientes parámetros:

R(a,b,c)	S(c,d)
N(R) = 5000	N(S)=200
V(R,a)=5000	
V(R,b)=3000	
V(R,c)=5	V(S,c)=5
	V(S,d)=40
Size(a)=20	
Size(b)=60	
Size(c)=20	Size(c)=20
	Size(d)=40

   Teniendo en cuenta que el tamaño de bloque es de 2KB, que la cabecera es de 20B y que en memoria sólo cabe un bloque, determina el número de operaciones de E/S que supondría la ejecución de la consulta:

$$\Pi_{a,d}(R\ JOIN\ S)$$

2. Explica qué elementos del nivel interno de Oracle se crearían y cuánto ocuparía la tabla recién creada si se ejecuta la consulta:

```
create table prueba(nombre varchar2(40) primary key, DNI
varchar2(8)) tablespace users
storage (initial 40k next 20k maxextents 10);
```

3. Se dispone de una relación R(a, b) donde a es la clave de valores únicos por la que se mantiene ordenado el archivo y b es un atributo con valores duplicados, además se tiene B = 4096B, C=10B, P= 10B, N (R) = 1000, V (R, b) = 200, size(a) = 10B, size(b) = 40B. Se montan dos índices I_A e I_B, uno por cada atributo. Calcula el tamańo en bloques de cada índice.

4. Propón dos planes lógicos para la siguiente consulta e indica de qué dependerá que se escoja uno u otro de sus planes físicos asociados:

$$\Pi_C(\sigma_{A=a \wedge B=b}(R))$$

5. Indica si son ciertas o falsas las siguientes afirmaciones y, brevemente, explica por qué:

   a) El tiempo que se tarda en reorganizar un fichero ASI (Archivos Secuenciales Indexados) denso depende únicamente del tiempo necesario para reordenar el fichero de desbordamiento y de reescribir ordenadamente el nuevo fichero maestro.

   b) Para consultas por rango, los ASI (Archivos Secuenciales Indexados) son más adecuados.

Soluciones

Pregunta 1

Establecemos que el número de operaciones de E/S coincide con el número de bloques que es necesario leer y escribir al disco, tanto de las relaciones originales como de las relaciones intermedias, necesarias para el cálculo del resultado.

Número de bloques en R:

L_R = 20 B + 60 B + 20 B = 100 B

Bfr_R = parte entera ((B – C) / L_R) = parte entera ((2048 B – 20 B) / 100 B) = 20 registros / bloque

Hay que leer *redondeo hacia arriba* (5000 / 20) = 250 bloques para cargar R

Para que la reunión natural sea eficiente, R debería estar ordenado por el atributo común, lo que implica $250 \cdot \log_2 250 = 1992$ bloques leídos

Número de bloques en S:

L_S = 20 B + 40 B = 60 B

Bfr_S = parte entera ((B – C) / L_S) = parte entera ((2048 B – 20 B) / 60 B) = 33 registros / bloque

Hay que leer *redondeo hacia arriba* (200 / 33) = 7 bloques para cargar S

Número de bloques en R JOIN S:

L_{JOIN} = 100 B + 60 B – 20 B = 140 B

Bfr_{JOIN} = parte entera ((B – C) / L_S) = parte entera ((2048 B – 20 B) / 140 B) = 14 registros / bloque

Número de tuplas del JOIN: $N_{JOIN} = N_R \cdot N_S / \max \{V(R,c), V(S,c)\} = 5000 \cdot 200 / 5 = 200000$ registros

Hay que escribir *redondeo hacia arriba* (200000 / 14) = 14286 bloques del resultado del JOIN

Hay que cargar 14286 bloques del resultado del JOIN para resolver la proyección

Número de bloques de la proyección

$L_{PROYECCION}$ = 20 B + 40 B = 60 B

$Bfr_{PROYECCION}$ = parte entera ((B – C) / $L_{PROYECCION}$) = parte entera ((2048 – 20) / 60 = 33 registros / bloque

Hay que escribir *redondeo hacia arriba* (200000 / 33) = 6061 bloques

En resumen, hay que leer y escribir:

- leer 1992 bloques de R para ordenar los 250 bloques,
- escribir los 250 bloques de R ordenados,
- leer 250 bloques de R ordenado,
- leer 7 bloques de S,
- escribir 14286 bloques del JOIN,
- leer 14286 bloques del JOIN, y
- escribir 6061 bloques de la proyección

o sea, **37132 bloques** en total.

Pregunta 2

Durante la creación de la tabla se crean:

- una segmento,
- una extensión y
- un bloque.

Teniendo en cuenta que la tabla tiene una clave primaria, también se crearía un índice.

La extensión inicial y el segmento ocupan un total de 40 KB.

Pregunta 3

En primer lugar, calcularemos el tamaño en bloques del índice construido sobre el atributo a.

El tamaño de cada entrada del índice sería:

$$L(I_A)=10\,B+10\,B=20\,B$$

Para ver cuántos registros entran en un bloque del índice, calculamos el **factor de bloqueo**:

$$Bfr(I_A)=\text{parte entera}\left(\frac{B-C}{L(I_A)}\right)=\text{parte entera}\left(\frac{4096\,B-10\,B}{20\,B}\right)=204$$

El número de bloques necesario para almacenar las 1000 entradas de la relación en el índice se calcula:

$$B(I_A)=\text{redondeo hacia arriba}\left(\frac{N(R)}{Bfr(I_A)}\right)=\text{redondeo hacia arriba}\left(\frac{1000}{204}\right)=5$$

El tamaño del índice I_A es de 5 bloques.

Para calcular el tamaño del índice construido sobre el atributo b, necesitamos hacer algunos cálculos adicionales.

Dado que un índice no puede tener valores repetidos, puesto que la búsqueda binaria no funcionaría, hemos de construir otro tipo de índice. En concreto, podríamos usar una estructura de índice multinivel en el que, el primer nivel (raíz) no tuviera valores repetidos, y el segundo nivel tuviera valores repetidos.

Siendo así, cada valor distinto del atributo b sólo podría aparecer una vez en el primer nivel. Para calcular cuántos bloques hacen falta para el primer nivel, tenemos que tener en cuenta el número de entradas:

$$B(I_B{}^1)=\text{redondeo hacia arriba}\left(\frac{N(I_B)}{Bfr(I_B)}\right)=\text{redondeo hacia arriba}\left(\frac{V(R,b)}{Bfr(I_B)}\right)$$

$$Bfr(I_B)=\text{parte entera}\left(\frac{B-C}{L(I_B)}\right)=\text{parte entera}\left(\frac{4096\,B-10\,B}{L(I_B)}\right)$$

$$L(I_B)=40\,B+10\,B=50\,B$$

Por lo que:

$$Bfr(I_B)=\text{parte entera}\left(\frac{4086\,B}{50\,B}\right)=81$$

$$B(I_B{}^1) = \text{redondeo hacia arriba}\left(\frac{V(R,b)}{Bfr(I_B)}\right) = \text{redondeo hacia arriba}\left(\frac{200}{81}\right) = 3 \text{ bloques}$$

Véase que el número de entradas del índice de primer nivel sería el número de valores distintos para el atributo b en la relación R, y que la estructura de todos los índices es exactamente igual, sean de un sólo nivel o multinivel, por lo que el cálculo del factor de bloqueo para el índice I_B es igual que el calculado para el índice I_A.

Con respecto al segundo nivel del índice, cada bloque contendrá entradas con el mismo valor de clave que se recorrerán secuencialmente. En ese caso, habría que calcular cuántos bloques se estiman necesarios para contener las 1000 entradas distintas de la tabla agrupadas en tantos grupos como valores distintos haya del atributo b.

El número de bloques estimado para contener los registros con valores iguales de b se calcula:

$$\text{redondeo hacia arriba}\left(\frac{N(I^2{}_B)}{Bfr(I_B)}\right) = \text{redondeo hacia arriba}\left(\frac{\frac{N(R)}{V(R,b)}}{Bfr(I_B)}\right) =$$

$$= \text{redondeo hacia arriba}\left(\frac{\frac{1000}{200}}{81}\right) = \text{redondeo hacia arriba}\left(\frac{1000}{200 \cdot 81}\right) = 1$$

De este modo, para cada uno de los 200 valores de b, se estiman 5 registros que caben en un bloque. De ese modo, cada valor distinto tendría 1 bloque en el segundo nivel del índice, lo que hacen:

$$B(I^2{}_B) = 1 \text{ bloque / valor} \cdot 200 \text{ valores} = 200 \text{ bloques}$$

De este modo, el cálculo total de bloques para el índice I_B es:

$$B(I_B) = B(I^1{}_B) + B(I^2{}_B) = 3 + 200 = 203 \text{ bloques}$$

El tamaño del índice I_B es de 203 bloques.

Pregunta 4

Alternativa 1

$$\begin{array}{c}\Pi_{C}\\ |\\ \sigma_{A=a}\\ |\\ \sigma_{B=b}\\ |\\ R\end{array}$$

Plan a

$$\begin{array}{c}\Pi_{C}\\ |\\ \sigma_{B=b}\\ |\\ \sigma_{A=a}\\ |\\ R\end{array}$$

Plan b

Las relaciones iniciales de ambos planes (R) son las mismas (mismo tamaño, mismos bloques, mismas tuplas). Lo mismo ocurre con las segundas selecciones en cada plan y con las relaciones resultantes entre sí. La diferencia estará en la primera selección y el criterio de elección dependerá del tamaño de la relación intermedia resultante de la primera selección en cada plan.

Siendo así, para cada relación conocemos *V(R,A)* y *V(R,B)*. El número de tuplas en ambas es de:

$$N_1=\frac{N(R)}{V(R,A)} \text{ y } N_2=\frac{N(R)}{V(R,B)}$$

Se escogerá el primer plan (a) sobre el segundo (b) si el tamaño de la primera selección en el plan a, en número de tuplas, es menor que el de la primera selección en el plan b:

$$\frac{N(R)}{V(R,A)}<\frac{N(R)}{V(R,B)}\Rightarrow\frac{1}{V(R,A)}<\frac{1}{V(R,B)}\Rightarrow V(R,B)<V(R,A)$$

es decir, se escoge el primer plan si la variabilidad del atributo B en la relación es menor que la variabilidad del atributo A.

Alternativa 2

$$\begin{array}{c}\Pi_{C}\\ |\\ \sigma_{A=a}\\ |\\ \Pi_{A,C}\\ |\\ \sigma_{B=b}\\ |\\ \Pi_{A,B,C}\\ |\\ R\end{array}$$

Plan a

$$\begin{array}{c}\Pi_{C}\\ |\\ \sigma_{B=b}\\ |\\ \Pi_{B,C}\\ |\\ \sigma_{A=a}\\ |\\ \Pi_{A,B,C}\\ |\\ R\end{array}$$

Plan b

Las relaciones iniciales de ambos planes (R) son las mismas (mismo tamaño, mismos bloques, mismas tuplas). Se aplica la misma consideración a las relaciones resultantes y a la primera proyección en cada plan.

En cualquiera de los casos, se escogerá el plan que reduzca el número de bloques en mayor número antes que el otro.

El número de tuplas en la primera selección del plan (a) y del plan (b) es de:

$$N_1=\frac{N(R)}{V(R,A)} \text{ y } N_2=\frac{N(R)}{V(R,B)}$$

Dado que ambas relaciones intermedias tienen el mismo esquema, tendrán el mismo tamaño en las tuplas L_1 y, por tanto, el factor de bloqueo será exactamente el mismo Bfr_1, siendo el número de bloques:

$$Bloques_1=\frac{N(R)}{V(R,A)\cdot Bfr_1} \text{ y } Bloques_2=\frac{N(R)}{V(R,B)\cdot Bfr_1}$$

Si comparamos, el número de bloques sólo dependerá de la variabilidad de los atributos, escogiendo el plan (a) si la variabilidad de B es menor (igual que en la alternativa

1).

La segunda proyección es asunto distinto, ya que los resultados de la relación intermedia en cada plan no tiene por qué tener el mismo tamaño en número de bloques.

$$Bfr_{a,2}=\frac{B-C}{V_A+V_C} \text{ y } Bfr_{b,2}=\frac{B-C}{V_B+V_C}$$

El tamaño en bloques de cada relación será de:

$$Bloques_{a,2}=\frac{N(R)}{V(R,A)\cdot\frac{B-C}{V_A+V_C}} \text{ y } Bloques_{b,2}=\frac{N(R)}{V(R,B)\cdot\frac{B-C}{V_B+V_C}}$$

Se escogerá el plan (a) sobre el (b) si el número de bloques de dicho plan es menor:

$$\frac{N(R)}{V(R,A)\cdot\frac{B-C}{V_A+V_C}}<\frac{N(R)}{V(R,B)\cdot\frac{B-C}{V_B+V_C}}\Rightarrow\frac{1}{V(R,A)\cdot\frac{B-C}{V_A+V_C}}<\frac{1}{V(R,B)\cdot\frac{B-C}{V_B+V_C}}\Rightarrow$$

$$\frac{V_A+V_C}{V(R,A)\cdot(B-C)}<\frac{V_B+V_C}{V(R,B)\cdot(B-C)}\Rightarrow\frac{V_A+V_C}{V(R,A)}<\frac{V_B+V_C}{V(R,B)}$$

Pregunta 5

a) **Falso**, porque depende del tiempo que tardemos en reordenar el fichero de desbordamiento, el tiempo que se tarda en reescribir ordenadamente el nuevo fichero maestro (mezclando el fichero maestro antiguo con el de desbordamiento reordenado) y el tiempo que tardamos en regenerar el índice.

b) **Verdadero**, pero esto sólo ocurre cuando se consulta por el mismo atributo del índice. Si no, es altamente ineficiente.