Parcialito Procesamiento de Consultas

lunes, 19 de junio de 2023 11:25

Formulas:

Selección:

File Scan: $B(R)$

Idx Scan:
- Primario: $H(I) + 1$
- Secundario: $H(I) + \left\lceil \frac{n(R)}{V(A_i,R)} \right\rceil$
- Clustering: $H(I) + \left\lceil \frac{B(R)}{V(A_i,R)} \right\rceil$

Join

Loops Anidados: $B(R) + \left\lceil \frac{B(R)}{M_{utiles}-1} \right\rceil * B(S)$ $\quad B(R) \leq B(S)$

Unico Loop: $B(R) + n(R) * Costo\_sel\_idx(S)$ $\quad \exists I(A_i,S)$

Hash Grace: $3\,(B(R)+B(S))$ con suf memoria para particiones

EJERCICIO 1 - Dado el siguiente esquema sobre profesores de una facultad:

- profesores(legajo, nombre, apellido, genero, titulo)

Se pide:

1. Analice cuál es el método de acceso más eficiente para resolver la siguiente consulta y estime su costo (en bloques accedidos):

$$\sigma_{titulo=\text{“}LICENCIADO\text{”} \wedge genero=\text{“}M\text{”}}(profesores)$$

2. Estime la cantidad de tuplas devueltas por dicha selección.

Información que se posee:

- n(profesores) = 500 y F(profesores) = 25.
- V(titulo, profesores) = 10 y V(genero, profesores) = 2.
- La tabla profesores tiene un índice llamado prtit por título con Height(prtit) = 2. Este índice **no** es de clustering.
- Los valores de las columnas titulo y genero se almacenan siempre en mayúsculas.

$n(R) = 500 \Rightarrow B(R) = 20$
$F(R) = 25$

$V(t,R) = 10$
$V(g,R) = 2$

$\exists I(t,R)\,,\ H(I(t,R)) = 2$

→ Se puede usar indice si es necesario

1) Full Scan: Condiciones se pueden aplicar al mismo tiempo ⇒ Se escanea una sola vez

Costo: $B(R) = 20$

Utilizando indice por titulo

Costo de usar indice (no clustering): $H(I(t,R)) + \left\lceil \frac{n(R)}{V(t,R)} \right\rceil = 2 + \left\lceil \frac{500}{10} \right\rceil = 2 + \underset{*}{50} = \underset{*}{52}$

La segunda condicion se puede aplicar a la salida de cada tupla en la seleccion con indice sin costo I/O

* No existen mas de 20 bloques, por lo que esta estimacion no aplica de esta manera
⇒ Se leeran a lo sumo los 20 bloques

⇒ Cost indice = 22

2) Dado que los atributos son independientes (o al menos se espera eso), asumiendo dist uniforme en ambas

$$n(\sigma) = \frac{n(R)}{V(t,R) \cdot V(g,R)} = \frac{500}{10 \cdot 2} = 25$$

EJERCICIO 2 - Dado el siguiente esquema que registra cuando un usuario le da “Me gusta” a una publicación:

- megusta(id_usuario, id_publicación, fecha_hora)

Se busca encontrar pares de usuarios a los que les guste una misma publicación. Se pide:

1. Analice cuál es el método de acceso más eficiente para resolver la siguiente consulta y

$n(R) = 100 \times 10^6 \Rightarrow B(R) = 100\,000$
$F(R) = 100 \times 10^1$

$V(u,R) = 50 \times 10^3$
$V(i,R) = 10 \times 10^6$

EJERCICIO 2 - Dado el siguiente esquema que registra cuando un usuario le da "Me gusta" a una publicación:

- megusta(id_usuario, id_publicación, fecha_hora)

Se busca encontrar pares de usuarios a los que les guste una misma publicación. Se pide:

1. Analice cuál es el método de acceso más eficiente para resolver la siguiente consulta y estime su costo (en bloques accedidos):

$$megusta \bowtie_{\substack{id\_usuario \neq id\_usuario' \wedge \\ id\_publicacion = id\_publicacion'}} megusta'$$

2. Estime la cantidad de tuplas devueltas por dicha selección.

Información que se posee:

- n(megusta) = 100,000,000 y F(megusta) = 1,000
- V(id_usuario, megusta) = 50,000 y V(id_publicación, megusta) = 10,000,000
- No se cuenta con índices y se dispone de M = 1,001 bloques de memoria disponibles.

$n(R) = 100 \times 10^6$ $\quad F(R) = 100 \times 10^1$ $\Rightarrow B(R) = 100\,000$

$V(u,R) = 50 \times 10^3$
$V(i,R) = 10 \times 10^6$

$M = 1001$

No indices

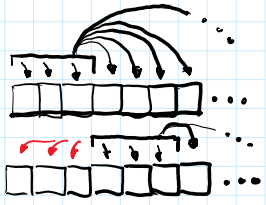
1 para guardar resultados

Usando loops anidados las condiciones se pueden verificar al momento de evaluacion

Se puede cargar 999 bloques de la relacion por vez, dejando B(R)-999 bloques para comparar uno por vez contra esos 999

Se debe leer todos los bloques al menos una vez

$B(R)$



Por cada 999 se deben cargar

$$N = \left\lceil \frac{B(R)}{M_{utiles} - 1} \right\rceil$$

$$\sum_{i=1}^{N} B(R) - \left(i * (M_{utiles} - 1)\right)$$

6 Comparaciones ya realizadas

Luego el costo sera

$$B(R) + \underbrace{\left(\sum_{i=1}^{N} B(R) - \left(i * (M_u - 1)\right)\right)}_{N * B(R) - \frac{N(N-1)}{2} \cdot (M_u - 1)}$$

$$B(R) + N * \left(B(R) - \frac{(N-1)}{2} \cdot (M_u - 1)\right)$$

$X - Y + X - 2Y$

$N \cdot B(R) - \frac{N(N-1)}{2} \cdot Y$

$N = 101$

$100\,000 + 101 * \left(100.000 - 50.999\right) = 5.054.151$

Si se hubiese tratado como dos relaciones diferentes

$100\,000 + 101 * 100\,000 = 10.200.000$

Con junta Hash Grace

$$\left\lceil \frac{V(id,R)}{K} \right\rceil$$

Con junta Hash Grace

Con 500 particiones ⇒ $\frac{100000}{500}$ = 200 bloques por particion

Particionando por id_publicación

$\left\lceil \frac{10.000.000}{500} \right\rceil = 200000$

$\left\lceil \frac{V(id,R)}{K} \right\rceil$

200K dif valores id-publicación por partición

Costo Hash

$2 * B(R)$

Max cant tuplas para una particion

$M_U * F(R) = 1000 * 1000 = 10^6$

↳ Permite un Hasheo aprox uniforme

Costo comparacion

Se comparan solo bloques de la misma particion para ver si cumplen con ambas condiciones

↳ B(R)

El costo total sera

$3 * B(R) = 300000$

Hash Grace esta realizando un "ordenamiento" implicito

Utilizando Sort Merge

Costo Sort

$2 * B(R) \cdot \lceil \log_{M_U}(B(R)) \rceil = 2 * 100000 \cdot \lceil \log_{1000} 100000 \rceil = 4 \times 100000 = 400000$

Costo Merge

$B(R) = 10000$

Costo total

$2 * B(R) \cdot \lceil \log_{M_U}(B(R)) \rceil + B(R) = 500000$

El mejor metodo sigue siendo HASH-GRACE

2) Suponiendo distribucion uniforme e independencia entre los atributos

Por cada combinacion U_1, U_2 quiero encontrar las publicaciones en comun

- Cada usuario dio me gusta a $\frac{n(R)}{V(u,R)}$ publicaciones
- Cada publicacion tiene $\frac{n(R)}{V(i,R)}$ me gusta
- Cada usuario coincidira con aprox $\frac{n(R)}{V(i,R)} \cdot \frac{n(R)}{V(u,R)}$ usuarios
- Esto contado por cada usuario $\frac{V(u,R) \cdot n(R) \cdot n(R)}{\quad}$

$V(i,R)$ $V(u,R)$

- Esto contado por cada usuario $\dfrac{V(u,R) \cdot n(R) \cdot n(R)}{V(i,R) \cdot V(u,R)}$

$$n\left(R \bowtie R'\right) = \frac{n(R)^2}{V(i,R)} = \frac{10^{16}}{10^7} = 10^9$$