OPTIMIZACION DE CONSULTAS A BASES DE DATOS RELACIONALES

1. CONTENIDO DEL MARCO TEORICO
2. OPTIMIZACION ALGEBRAICA DE CONSULTAS
3. COSTO DEL INPUT / OUTPUT DEL PRODUCTO CARTESIANO Y JUNTA NATURAL
4. TUNING

1. CONTENIDO DEL MARCO TEORICO.

- Procesamiento de consultas.
  Propiedades de los operadores algebraicos. Optimización algebraica. Información del catálogo para la estimación del costo. Medidas del costo de una consulta. Estimación del costo de procesamiento de consultas: Costo de cada operador algebraico; cálculo del costo del input y output del producto cartesiano. Estrategias para procesar una consulta contando con índices.

- Tuning de bases de datos relacionales.
  Impacto de un buen diseño físico. Determinación del diseño físico según la Esperanza de accesos. Reglas empíricas de indexación. Importancia del conocimiento sobre el funcionamiento del optimizador de consultas del RDBMS. Mejorando el código SQL y su impacto en la performance. Casos reales.

2. OPTIMIZACION ALGEBRAICA DE CONSULTAS.

EJERCICIO 1

Optimizar la siguiente expresión algebraica utilizando el árbol de consulta:

$$\Pi_{ART.desc} \quad \sigma_{F} \quad (\quad (PEDIDO \ X \ CLIENTE \ X \ ART \ ))$$

donde F es:

PEDIDO.nc = CLIENTE.nc
and PEDIDO.na = ART.na
and PEDIDO.q = 200
and ART.color = "rojo"
and ART. Talla = "2"
and CLIENTE.cd = "Rosario"

EJERCICIO 2

Sea D={Q,R,S} una descomposición con los esquemas: Q (A,B), R(B,C) y S(AC). q(Q), r(R), s(S) son una proyección de una instancia definida sobre ABC, usar el método de optimización de consultas para hallar expresiones equivalentes a las siguientes co-juntas:

a) Π_B (q |X| r |X| s)

b) Π_{AC} (q |X| r |X| s)

c) Π_{AC} (($\sigma_{A=c1}$ q) |X| ($\sigma_{C=c2}$ r) |X| s)

EJERCICIO 3

Especifique los pasos generales para optimizar la siguiente expresión:

Π_B ($\sigma_{A=a}$ ($\sigma_{D=d}$ (q) |X| (r - Π_{BDF} (r|X|s))))

donde: q(Q); r(R); s(S) y Q(ABD); R(BDF); S(FG) son relaciones con sus respectivos esquemas.

3. COSTO DEL INPUT / OUTPUT DEL PRODUCTO CARTESIANO Y JUNTA NATURAL.

EJERCICIO 4

Sea n_1 y n_2 el número de registros de r_1 y r_2 respectivamente, y sea p el número de bloques que pueden alojarse en memoria principal. Deduzca una fórmula que de el número total de accesos para efectuar : r_1 X r_2. Estudie en términos de tiempo de ejecución la diferencia que existe entre r_1 X r_2 y r_2 X r_1.

EJERCICIO 5

En un centro de cómputos se quiere computar el producto cartesiano de r x s.

Sea Br la cantidad de bloques que deben ser accedidos para leer el archivo r y Lr es la cantidad de by necesarios para almacenar una tupla de r ; análogamente es Bs y Ls.
Sea b la cantidad de by que hay en un bloque.
Sea M la cantidad de bloques que entran en memoria principal al mismo tiempo.

Se pide calcular :

5.1) El costo de leer el input.

5.2)

5.2.1) La cantidad aproximada de tuplas de r y de s (|r| y |s|) en función de los datos dados.

5.2.2) El costo para grabar el output en función de: Br, Bs, r y s . (ayuda : calcule la cantidad aproximada de bloques necesarios para almacenar el output.)

5.2.3) El costo total del proceso en función de: Br, Bs, r y s.

5.3) Como el proceso era muy lento el jefe de desarrollo pidió invertir dinero en una ampliación de la memoria principal. Una vez ya instalada y luego de realizar el mismo proceso exclamó: no puede ser ! cuadriplique la memoria principal y solo mejoré el tiempo en un 2% !. Explíquele al jefe que es lo que sucedió.

EJERCICIO 6

Se dice que el tamaño del output para una junta natural que se realiza por el atributo A que no es CC de ninguno de los dos esquemas esta acotado por :

$$\frac{|r_1| \times |r_2|}{D(A, r_1)}$$

Se pide : a) Explique bajo que hipótesis lo anterior es verdadero

4. TUNING.

4.1) Defina los siguientes términos :

Índice primario
Índice secundario
Índice denso
Índice escaso o poco denso
Índice múltiple
Clave de búsqueda
Cluster
Índice cluster

4.2) En general, es posible tener dos índices cluster en la misma relación para diferentes claves de búsqueda? Justifique.

4.3) Puesto que los índices agilizan el procesamiento de las consultas, por qué no deberían mantenerse en varias claves de búsqueda? Listar tantas razones como sea posible.

4.4) Si se utiliza una Organización física Relativa usando la técnica de Hashing, en una clave de búsqueda para la cual es probable que se hagan consultas de rangos, qué propiedad deberá tener la función de asociación?

4.5) Para cada una de las siguientes proposiciones decir si es, en general, verdadera o falsa justificando su respuesta :

4.5.1) Indexar las columnas que son usadas frecuentemente en la cláusula where.

4.5.2) Indexar las columnas cuyos valores max y min se seleccionan frecuentemente.

4.5.3) Indexar las columnas que tengan un alto poder selectivo (la selectividad es alta si son pocas las filas que tienen el mismo valor en la columna clave).

4.5.4) Busque tantos casos particulares como pueda para contrariar sus respuestas anteriores.

4.6) Sea un archivo cuyos registros tienen tres campos : A,B,C. Hay 10 valores posibles para A, 100 valores posibles para B, y 20 valores posibles para C. Suponga que TODOS los registros posibles se encuentran en el archivo. Suponga que todas las consultas especifican valores para exactamente dos campos: La probabilidad de que una consulta especifique A y B es de 0.8; la probabilidad de que especifique A y C es 0.15; y la probabilidad de que especifique B y C es 0.05. Suponga que puede crear un solo índice múltiple. Qué campos deben elegirse ?

4.7) Sean los esquemas de relación R(A,B,C,D) y S(E,F,G,A) y los siguientes datos para las respectivas instancias :

r(R) :	s(S) :
|r| = 30.000 ;	|r| = 10.000 ;
FB = 3 ;	FB= 10 ;
V(B,r) = 10000 ;	V(F, s) = 2000 ;
	V(A,s) = 8000 ;

Suponga que se realizan solo las tres consultas que a continuación se detallan y que c/u tiene la siguiente probabilidad : 0.55 , 0.15 y 0.30 respectivamente.

4.7.1)
```
SELECT A , B , C , D
FROM R , S
WHERE R.A = S.A ;
```

4.7.2)
```
SELECT *
FROM R
WHERE R.B = “x” ;
```

4.7.3)
```
SELECT *
FROM S
WHERE S.F = “Y” ;
```

Calcule justificando su respuesta lo siguiente :

a) La esperanza de accesos en un diseño sin índices ni clusters. Debe tener en cuenta en cada caso, el costo del input y del output de cada consulta.

b) El mejor diseño físico y su esperanza de accesos si se quiere emplear solamente un índice.

c) Ídem a b) pero usando solamente dos índices.

d) El mejor diseño físico y su esperanza de accesos usando no más de dos índices.

En todos los casos, si es necesario, debe aclarar las suposiciones adicionales.

4.8) Sea el esquema de relación R(A,B,C,D,E) y los siguientes datos para una instancia :

r(R):
|r| = 30.000;
V(A,r) = 2
V(B,r) = 30.000;
V(C,r) = 10.000;
V(D,r) = 1.000;
V(E,r) = 2;
FB=5;

Suponga que se realizan solo las cinco consultas que a continuación se detallan y que c/u tiene una p(i) probabilidad de ser utilizada.

4.8.1)
```
SELECT *
FROM R
WHERE R.A = “X” ;
```

4.8.2)
```
SELECT *
FROM R
WHERE R.B < 60.000 ;
```

4.8.3)
```
SELECT A , D
FROM R
WHERE R.C = “Y” ;
```

4.8.4)
```
SELECT *
FROM R
WHERE R.D = “Z” ;
```

4.8.5)
```
SELECT *
FROM R
WHERE R.B = ( SELECT MAX(B) FROM R) ;
```

Asigne usted una probabilidad a cada consulta (que no sean iguales) y calcule justificando su respuesta lo siguiente :

a) La esperanza de accesos en un diseño sin índices.

b) La menor esperanza de accesos con un diseño que emplee un solo índice.

c) Ídem a b) pero usando solamente dos índices.

d) El mejor diseño físico que Ud. considere.

En todos los casos, si es necesario, debe aclarar las suposiciones adicionales.

4.9) Dada la siguiente consulta SQL :

```
SELECT NOMBRE , NRO_DEP
FROM departamento
WHERE NRO_DEP NOT IN ( SELECT NRO_DEP FROM empleado ) ;
```

Existe un índice para el atributo NRO_DEP de empleado pero el optimizador de consultas del RDBMS que utiliza (ORACLE x) no es demasiado "listo" y no utiliza índices si no se nombra en la cláusula where la columna(s) clave(s) correspondiente(s) al índice.

Se pide :

a) El costo del input.

b) Cómo modificaría la consulta para “ayudar” al optimizador a mejorar la performance y cuál sería el nuevo costo del input ?

4.10) Dada la siguiente consulta SQL :

```
SELECT DISTINCT NF
FROM ped ped1
WHERE NOT EXISTS ( SELECT *
                   FROM ped ped2
                   WHERE NOT EXISTS ( SELECT *
                                      FROM ped ped3
                                      WHERE ped3.NP = ped2.NP
                                            AND
                                      ped3.NF = ped1.NF ) ) ;
```

Suponiendo que todo índice entra en MP y que existen un índice por el atributo np.

Se pide :

a) El costo del input (para M = 2).

b) Puede disminuir el costo anterior con un diseño físico más adecuado a dicha consulta ? Si contesta afirmativamente indique cual es el diseño físico correspondiente y el nuevo costo del input.

c) Qué pasará si reemplazamos el primer SELECT ..... FROM ..... por SELECT NF FROM fab y cambiamos ped1.NF por fab.NF en la última línea ?

4.11) Se ha creado un índice múltiple (secundario y denso) de la siguiente manera :

CREATE INDEX PEPE ON empleado (APELLIDO , CIUDAD , PROVINCIA) ;

Encuentre una estrategia inteligente para cada una de las siguientes consultas dando en cada caso su costo :

a) Listar el apellido y ciudad de todos los empleados.

b) Listar a todos los empleados que viven en Mar del Plata.

c) Listar las ciudades donde viven los empleados de apellido González.