Assignment 3 Solution

Joining Techniques

Instructions : You will have to take assumptions for block size (B), record size (R), number of records (r), available memory (K), indexing column, index type (clustered/non-clustered), and all remaining factors required in calculations.

Note: The following question is a little different from assignment. But it will give you a clear concept.

Question: Consider two tables A and B which have to be joined. Calculate the cost of joining the two tables on their common attribute. You will have to consider the following cases and have to calculate cost for all these cases:

1. When the table sizes are almost similar. Let's say 1 row of table A joins with only 1 row of table B.

- When RAM size is not sufficient.
- When RAM size sufficient for both tables.

2. When the sizes of both tables are significantly different. Let's say 1 row of table A joins with 10 rows of table B.

- When RAM size is not sufficient.
- When RAM size sufficient for both tables.

Your solutions should include costs of I/Os for **Naïve nested loop join (or Block NLJ/Index NLJ/clustered index NLJ), Sort merge join (or merge join), and in-memory hash join**.

Let the query be

Select *

From A join B

Where A.x= B.x

Let x be the joining column.

B= Block size = 4k

R= Record Size = 256 Bytes

rA = number of records of table A = 64000 rB = number of records of table B = 64000

bfr = B/R = (4* 1024)/256 = 16

bA = data blocks of A = rA /bfr = 64000/16 = 4000

bB = data blocks of B = rB /bfr = 64000/16 = 4000

Let Record size of index=Ri = 64 Byte

ri = number of records of index table=64000

bfri=(4*1024)/64 =64

bi = ri/bfri= (64000/64) = 1000

When RAM is not sufficient

Naive Nested loop

Cost =A+Qualifying records * Blocks of B

(As there is one to one mapping of rows)

Cost = 4000 + (64000*4000)

Block Nested loop join

Cost =A+Qualifying blocks of A * Blocks of B

(As there is one to one mapping of rows)

Cost = 4000 + (4000*4000)

Index nested loop join:

Lets say we have non –clustered index in table B on column x.

We will not read all the blocks of inner table for each qualifying row of outer table.

Number of blocks read from inner table for each qualifying row of outer table = Number of rows B that join with 1 row of A. (We will add index access cost too).

Clustered Index nested loop join:

Lets say we have clustered index in table B on column x.

We will not read all the blocks of inner table for each qualifying row of outer table.

Number of blocks read from inner table for each qualifying row of outer table = Number of qualifying blocks of B that join with 1 row of A. (We will add index access cost too).

Merge JOIN:

If tables are pre sorted

O(A+B)=4000+4000=8000

If Table B is sorted

O(A*log(A/k))+O(A+B) , where A and B are number of blocks of table A and B and k is the number of blocks of available memory.

If no table is pre sorted

O(Alog(A/k))+ O(Blog(B/k))+O(A+K)

Hash Join:

As memory Is not sufficient,no table fits in memory , cost will be

O(A log(A/k))+ O(B log(A/k))+O(A+B) , In general A is the number of blocks of smaller table

RAM is sufficient:

Sort Mege Join

Cost = O(A+B) whether tables are sorted or not

Nested Loop Join:

Cost = O(A+B)

Hash Join:

Cost = O(A+B)