CENG 352 - Database Management Systems
2023-2
Written Homework 2

Anıl Eren Göçer
e2448397@ceng.metu.edu.tr

May 5, 2024

Q1.

1.

Selectivity factor is $X = \dfrac{1}{V(T,D)} = \dfrac{1}{2000}$

Therefore, it returns $X.T(T) = \dfrac{1}{2000}.200000 = 100$ tuples.

10 tuples occupy 1 block $\longrightarrow B(T) = \dfrac{1}{10}.T(T) = \dfrac{1}{10}.100 = 10$ blocks.

2.

$$T(R \bowtie_B S) = T(R).T(S).\frac{1}{max(V(R,B), V(S,B))} = 2000.10000.\frac{1}{max(200, 2000)} = 10000$$

We found that $T(R \bowtie_B S) = 10000$. That is, the number of tuples it returns is 10000.

Observe that 2000 tuples of R fits into 200 pages. So, one tuple of R fits into $\dfrac{1}{10}$ page. Since attributes of R, which are A and B, are integers, one integer fits into $\dfrac{1}{20}$ pages.

Similarly, observe that 10000 tuples of S fits into 1000 pages. So, one tuple of S fits into $\dfrac{1}{10}$ page.

Therefore, we confirm that two attributes of S, namely B and C, are integers and fits into $\dfrac{1}{10}$ pages. Hence, one integer fits in $\dfrac{1}{10}$ pages.

One tuple of $R \bowtie_B S$ has three attributes, A, B and C. So, one tuple of this join occupies $\dfrac{3}{20}$ page (block).

Then, $B(R \bowtie_B S) = \dfrac{3}{20}.T(R \bowtie_B S) = \dfrac{3}{20}.10000 = 1500$ blocks returned.

3.

$$B(R) + \frac{B(R).B(S)}{M-2} = 200 + \frac{200.1000}{40} = 5200 \text{ I/O's.}$$

4.

$$M_1 = \frac{B(R)}{M} = \frac{200}{42} = 4.76 \longrightarrow 5 \text{ runs for R}$$

$$M_2 = \frac{B(S)}{M} = \frac{1000}{42} = 23.81 \longrightarrow 24 \text{ runs for S}$$

$M_1 + M_2 = 5 + 24 = 29 \leq 42 = M$

Also, $B(R) \leq M^2$ and $B(S) \leq M^2$

So, cost is $3B(R) + 3B(S) = 3.200 + 3.1000 = 3600$ I/O's.

5.

Condition to be verified is $min(B(R), B(S)) \leq M^2$

$min(200, 1000) \leq 1764 \longrightarrow 200 \leq 1764$

The condition is satisfied, so the cost is:

$3B(R) + 3B(S) = 3.200 + 3.1000 = 3600$ I/O's

6.

Note that the index on S.B is unclustered.

Cost: $B(R) + \dfrac{T(R).T(S)}{V(S,B)} = 200 + \dfrac{2000.10000}{2000} = 10200$ I/O's

7.

1st step: Apply BNLJ to $R \bowtie_B S$:

Read and join: $B(R) + \dfrac{B(R).B(S)}{M-2} = 200 + \dfrac{200.1000}{40} = 5200$ I/O's

This intermediate result (T1) occupy 1500 blocks (please refer to part 2 of this question).

Write T1 to disk: 1500 I/O's

For this step, in total: 6700 I/O's

2nd step: Apply selection (D =1500) to T:

Since we will do file scan, we need to read whole table. So, this requires B(T) = 20000 I/O's

The intermediate result (T2) will consist of 10 blocks (please refer to part 1 of this question)

We will need to write T2 back to disk. So, this requires 10 I/O's.

For this step, in total: 20010 I/O's.

3rd step: Apply BNLJ to $T1 \bowtie_C T2$:

Remember the schemas of T1 and T2 are T1(A,B,C) and T2(C,D).

Read and join: $B(T1) + \dfrac{B(T1).B(T2)}{M-2} = 1500 + \dfrac{1500.10}{40} = 1875$ I/O's

Now, we need to write this intermediate result (T3) to disk. Schema of T3 is T3(A,B,C,D).

Since T3 has 4 columns of type integer, one row of T3 occupies $\dfrac{1}{5}$ page (please refer to calculations in part 2 of this question).

Now, let's estimate the number of tuples in T3:

T(T1) = 10000 and T(T2) = 100 (please refer to part 1 and 2 for calculations)

$$T(T3) = T(T1).T(T2).\frac{1}{max(V(T1,C),V(T2,C))} = 10000.100.\frac{1}{max(V(S,C).1/2000, V(T,C).1/2000)}$$
$$= 10000.100.\frac{1}{max(10000.\frac{1}{2000}, 2000.\frac{1}{2000})} = 10000.100.\frac{1}{5} = 200,000$$

We estimated that T(T3) = 200,000 tuples. Since 1 tuple of T3 occupies $\dfrac{1}{5}$ block, we found $B(T3) = 40000$ blocks.

We need to write this to disk: 40000 I/O's

For this step, in total: 41875 I/O's

4th step: Apply file scan to T3 with projection

We need to read whole T3: 40000 I/O's. That is all for this step.

As a result: 6700 + 20010 + 41875 + 40000 = 108585 I/O's

Q2.

a)

Selectivity factor for the condition $A = 5678$ is $X_1 = \dfrac{1}{V(R,A)} = \dfrac{1}{10000}$.

Selectivity factor for the condition $D = 1234$ is $X_2 = \dfrac{1}{V(T,D)} = \dfrac{1}{100}$.

Selectivity factor for the condition $R.B = S.B$ is $X_3 = \dfrac{1}{max(V(R,B),V(S,B))} = \dfrac{1}{max(V(R,B),200000)}$

$V(R,B)$ can be $200000 = T(R)$ at maximum. So, $X_3 = \dfrac{1}{200000}$.

Selectivity factor for the condition $S.C = T.C$ is $X_4 = \dfrac{1}{max(V(S,C),V(T,C))} = \dfrac{1}{max(5000,V(T,C))}$

$V(T,C)$ was not given. So, $max(5000, V(T,C))$ can be 5000 at minimum and can be $10000 = T(T)$ at maximum. Therefore, I will take the average of them which is 7500. So, $X_4 = \dfrac{1}{7500}$.

Therefore, the size of the query (in terms of tuples) is

$$X_1.X_2.X_3.X_4.T(R).T(S).T(T) = \frac{1}{10000}.\frac{1}{100}.\frac{1}{200000}.\frac{1}{7500} = 13.33$$

By rounding up 13.33, the estimated size of the query is 14 tuples.

b)

1.

Apply index scan to $R(A,B)$ with the condition $A = 5678$ using the unclustered index.

Selectivity is $X_1 = \dfrac{1}{V(R,A)} = \dfrac{1}{10000}$.

$X_1.B(R) = \dfrac{1}{10000}.2000 = 0.2 \longrightarrow 1$ block is read. That is, 1 I/O is required for reading.

Let's call this intermediate result $T1$, and we do not need to write it to disk. It can be kept in memory.

Note that $T(T1) = T(R).X_1 = 200000.\dfrac{1}{10000} = 20$.

2.

Apply index nested loop join (INLJ) to $T1 \bowtie_B S(B,C)$.

We need to iterate over $T1$, for each tuple fetch corresponding tuple(s) from S.

Remember that the index on S.B is clustered.

So, the cost is $\dfrac{T(T1).B(S)}{V(S.B)} = \dfrac{20.1000000}{200000} = 100$ I/O's . (Notice that I did not include the term $B(T1) = 1$ since T1 is already in memory.)

Let's call this intermediate result T2 and note that

$$T(T2) = T(T1).T(S).X_2 \text{ where } X_2 = \frac{1}{V(S,B)} = \frac{1}{200000}$$
$$= 20.10000000.\frac{1}{200000} = 1000$$

Since $T2$ fits into memory, we do not need to write to disk, we can keep it in memory.

3.

Apply index nested loop join (INLJ) to $T2 \bowtie_C T$. We need to iterate over $T2$, for each tuple fetch corresponding tuple(s) from T.

Remember that the index on T.C is unclustered. So, the cost is

$$\frac{T(T2).T(T)}{V(T,C)} = \frac{1000.10000}{V(T,C)} \text{ I/O's}$$

$V(T,C)$ is not given, so I will use the rule of thumb given in the slides for the selectivity for joins which is assumed to be $\frac{1}{10}$. Therefore, the cost is

$$\frac{1000.10000}{10} = 1,000,000 \text{ I/O's}$$

4.

Applying on the fly selection for $D = 1234$ won't require any disk I/O.

Therefore, in total

1 + 100 + 1,000,000 = 1,000,101 IO's required.

c)

1.

Apply index scan to $R(A,B)$ with the condition $A = 5678$ using the unclustered index.

Selectivity is $X_1 = \frac{1}{V(R,A)} = \frac{1}{10000}$.

$X_1.B(R) = \frac{1}{10000}.2000 = 0.2 \longrightarrow 1$ block is read. That is, 1 I/O is required for reading.

Let's call this intermediate result $T1$, and we do not need to write it to disk. It can be kept in memory.

Note that $T(T1) = T(R).X_1 = 200000.\frac{1}{10000} = 20$.

2.

Apply hash join to $T_1 \bowtie_B S$. Since $B(T1) \leq M$, we can apply one-pass algorithm.

So, the cost is $B(T1) + B(S) = 1 + 1,000,000 = 1,000,001$ I/O's .

Let's call this intermediate result $T2$ and estimate its size:

$$T(T2) = \frac{T(T1).T(S)}{max(V(T1,B),V(S,B))} = \frac{T(T1).T(S)}{V(S,B)} = \frac{20.1,000,000}{200,000} = 1000 \text{ tuples}$$

And, $B(T2)$ is approximately 100 blocks.

3.

Apply index scan to $T(C,D)$ with condition $D = 1234$.

Selectivity factor is $X = \dfrac{1}{V(T,D)}$

Let's call this intermediate result T3.

$T(T3) = \dfrac{1}{V(T,D)}.T(T) = \dfrac{1}{100}.10000 = 100$

Since we use unclustered index to scan, 100 I/O's are required. Also, B(T3) is approximately 10 blocks.

4.

Apply hash join to $T2 \bowtie_C T3$

Since $B(T2) \leq M$, we can apply one pass algorithm.

The cost is: $B(T2) + B(T3) = 100 + 10 = 110$ I/O's
As a result, total cost is:

$$1 + 1,000,001 + 100 + 110 = 1,000,212 \text{ I/O's}$$