Database Systems, CSCI 4380-01
Homework # 8
Due Thursday May 5, 2011 at 2 pm

Answer the following questions. Turn in a single text or PDF file in the assignment drop box.

Question 1 [40 points]. You are given the statistics and the queries below.

(a) SELECT * FROM R WHERE A=1

(b) SELECT * FROM R WHERE B='foo'

(c) SELECT * FROM R WHERE B='foo' AND C=3

(d) SELECT * FROM R,S WHERE R.A=S.A

(e) SELECT * FROM R,S WHERE R.A=S.A AND B='foo'

TUPLES(R)=500,000	PAGES(R)=10,000
TUPLES(S)=1,000,000	PAGES(S) = 5,000

Attribute	VALUES	MIN	MAX
R.A	1,000	1	5,000
R.B	40	'aardvark'	'pringles'
R.C	100,000	1	100,000
S.A	800	1	800

(1) Find the total number of tuples satisfying each query.

(2) Given your answer above, find the total number of pages it would take the store the result of each query (if the result were to be stored on disk).

Question 2 [20 points]. You are given the schedule below.

$$r_1(X)\, r_3(Y)\, r_3(W)\, w_3(W)\, w_4(X)\, r_4(Y)\, w_4(Y)\, w_2(Y)\, r_1(Z)\, w_1(Z) commit_1\, commit_2\, commit_3\, commit_4$$

(1) Is this schedule serializable?

(2) Is this schedule possible under two phase locking (2PL)? Explain your answer assuming a single type of lock is used. Assume transactions can request locks at any point in the schedule as long as it is before they are needed. According to 2PL, they can be released at any time as long as a new lock is not obtained once a lock is released by a transaction.

(3) Is this schedule possible under strict two phase locking (Strict 2PL)?

Question 3 [30 points bonus]. You are given the following query plan for $R(A, B, D, E, F, H, I)$ and $S(A, C, G)$. Compute the overall cost of this query. Show your work clearly.

TUPLES(R)=500,000 PAGES(R)=40,000
TUPLES(S)=10,000,000 PAGES(S) = 400,000
Index I on R.D, height 3 with 6,000 leaf nodes

Attribute	VALUES	MIN	MAX
R.A	10,000	1	10,000
R.D	100	1	100
S.A	1,000,000	1	1,000,000
S.C	1,000	1	1,000

Assume a disk page and memory block is 1024 bytes, attributes A, B, C each span 12 bytes each.

Note that you need to figure out the size of the output of each operation in terms of the number of pages and then figure out the cost of the next operation in the pipeline as a function of this.

After a selection, you can assume the number of distinct values for all attributes remain the same (upper bounded by the number of tuples in the relation of course). For example, after the selection on $D = 1$, the number of distinct values of A is the same unless there are fewer than 10,000 tuples in the output, in which case, the number of distinct values is equal to the number of tuples in the output.

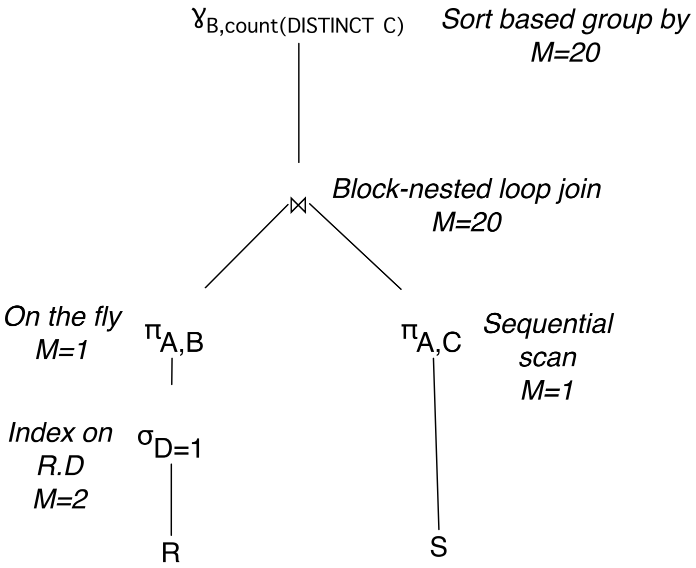

Figure 1: The query plan for question 3

Question 4 [30 points bonus]. Reduce the cost of attached query by either rewriting the query without changing its meaning or by introducing indices.

Note that your grade in this question will be proportional to the percentage improvement you achieve.

Use the latest database from Homework #6. This query is giving me a run time of 694 ms on this database.

Run `vacuum analyze` on your database before running the estimation routine and after creating an index to get reliable estimates of the cost of your query.

In your answers, describe what you have done: what indices you created and what the modified query is, and the modified query plan to show the improvement.