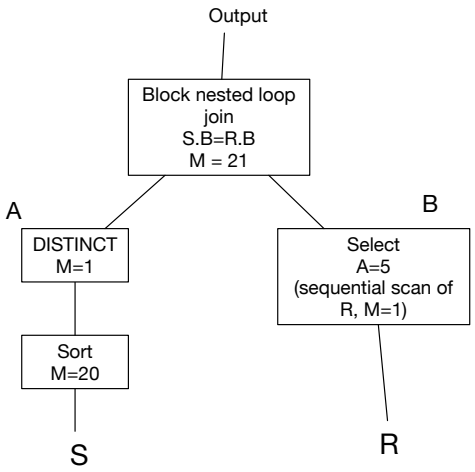
CSCI 4380 Spring 2018 Quiz 6 Answers



TUPLES(R)= 2,000,000
PAGES(R) = 8,000
VALUES(R.A)=40
VALUES(R.B)=20,000

TUPLES(S)= 800,000
PAGES(S) = 2,000
VALUES(S.B) = 16,000

Question (25 points). Suppose you are given the above query tree over two relations: `R(A,B,C,D)`, `S(B)`. Relation A is the result of duplicate removal of S and B is result of selection on R.

(a) Please estimate the following based on the given information.Show your work, write down calculations and explain with a single sentence why. You will not get full credit without an explanation.

Answer here.

Number of expected tuples in A:

16,000

Number of expected pages that will take to store A:

2,000* 16,000/800,000 = 40

Number of expected tuples in B:

2,000,000/40 = 50,000

Number of expected pages that will take to store B:

8,000/40 = 200

Number of expected tuples in A join B:

Potential ways to answer this:

50,000 (because at best each tuple in A will join with a single tuple in B)

or 20,000 unique values of B still after selection, so

50,000*16,000/(max(20000, 16000)) = 40,000

or, you can try to estimate which percentage of unique B values will survive after selection: 20000/40=500 and then use this:

50,000*16,000/(max(500,16000))= 50,000.

All reasonings will be accepted.

(b) What is the cost of this operation based on your estimates? Show your work explicitly below for (1) cost of sort, (2) cost of sequential scan, and (3) cost of join.

If the join requires multiple passes over R, we will accept the naive solution of reading R completely each time. In practice, if multiple passes are needed the result of the selection is written out to disk and read in iterations.

Answer here.

Sort of S:

Step 1: Cost: 4000, creates 2000/20=100 sorted groups

Step 2: Cost: 4000, reduces 100 groups to 5

Step 3: Cost: 2000: sort and output

Total = 10000

Distinct: Cost = 0 (0n the fly)

Select from R: Cost 8,000

Join: S after duplicate removal is 40 pages, so we need to read R 2 times

Cost: No cost for S (read from memory), read R 2 times:

2*8000=16,000

Total cost: 10,000 (sort and read S) + 16,000 (read R twice and do the selection each time) = 26,000

Advanced version of the solution (not required for the quiz): Do a sequential scan of R and write to output and read this in the second pass:

10000 (sort and read S) + 8,000 (read R) + 400 (write R once and read once for the second iteration of join) = 18,400.