Homework 5b

Assigned: 3/3/2020

Due: 3/9/20, 11:59PAM, electronically on canvas. Note this is <u>Monday before noon</u>. I had intended this to be due on Friday, together with homework 5a, but I also intended on having this posted on Box much earlier today.

Objectives:
  I. I/O (physical) cost of query operators, (more to come in hw6)
  II. [Logical] Query operations and cost wrt to size of results. To a first approximation, the applicable questions.walk you through each of the formulas stipulated by the text. Next homework will have problems that exercise the same calcultations, but on more complex queries.

Part A:
      Text problems 15.3.2, 15.3.3 a, 16.2.6 a,b

Part B:

  A. Use the following database schema and associated catalog data for all the problems in this homework. If it is not otherwise specified you should assume the query comprises a natural join.

schema	R(a,b,c)	S(a)	U(b)	V(c)	W(b,c)
blocks	B(R)= 1000	B(S)=5	B(U)= 20	B(V)= 10	B(W) = 100
rows	T(R)= 1,000,000	T(S)=1,000	T(U)=10,000	T(V)= 100	T(W) = 10,000
values	V(R,a) = 100	V(S,a)= 100			
	V(R,b) = 200		V(U,b)=400		V(W,b)=20
	V(R,c) = 50			V(c)= 100	V(W,c)= 10

I/O Cost, which is a function of the number of blocks read.
  1) In the linear cost model, what is the cost of reading the entire contents of relation R using a table scan?

  2) Assume R is 100% clustered. (i.e. all the blocks are contiguous on disk) In the affine cost model, what is the cost of reading the entire contents of relation R using a table scan?

  3) Assume c is the primary key of relation W.

      a. Using the linear cost model, what is the cost of reading the data pages of W, (i.e. ignore the B+tree nodes), for the following query?

      b. How many records do you estimate will be in the output?

    Select *
    From W

Where W.c = "Illinois"

4) Assume (c,b,a) is a compound primary key for relation R,

    a. Using the linear cost model, what is the cost of reading the data pages of R, (i.e. ignore the I/O for the B+tree nodes), for the following query?

    b. How many records do you estimate will be in the output?

    Select *
    From R
    Where R.a = 'Urbana' AND R.b = 'Champaign' AND R.c = 'Illinois'

5) Assume (c,b,a) is a compound primary key for relation R,

    a. Using the linear cost model, what is the cost of reading the data pages of R, (i.e. ignore the I/O for the B+tree nodes, but do assume you have the benefit of the primary index), for the following query?

    b. How many records do you estimate will be in the output?

    Select *
    From R
    Where R.b = 'Champaign' AND R.c = 'Illinois'

6) Assume that W has no useful index for the following question. How many rows do you estimate to be in the output of the following queries

    a. Select *
    Fr om W
    Where W.b = 'Champaign' **AND** W.c = 'Illinois'

    b. Select *
    From W
    Where W.b = 'Champaign' **OR** W.c = 'Illinois'

7) How many rows do you estimate to be in the output of the following queries? (Note, there is a change in data type. It does not effect the problem, or the answer, so you need not be concerned.)

    a. Select *
    From R
    Where W.b > 10' AND W.c = 'Illinois'

    b. Select *
    From W
    Where W.b > 10 **OR** W.c = 'Illinois'

Join Size Result: What is the size, in rows, of the following join queries. (Let |X| represent the join symbol)

8) S |X| U

9) V |X| W

10) R |X| W

11) $(\sigma_{c> 50} \ R)$ |X| U

12) $\underset{\text{R.b > W.b and R.c = W.c}}{R \quad |X| \quad W}$