Database Systems, CSCI 4380-01
Homework # 11
Due Monday December 9, 2019 at 11:59:59 PM

Homework Statement. This homework is worth 7% of your total grade. If you choose to skip it, Final Exam will be worth 7% more. This homework is for cost estimation and transactions (to be covered on Thursday 12/5).

SUBMISSION INSTRUCTIONS. Submit a single PDF file containing your solutions to Gradescope.

Question 1. What is the cost of following operations given below:

$PAGES(R) = 8,000 \ PAGES(S) = 100,000.$

(a) Sort R using $M = 100$ blocks.

(b) Sort S using $M = 100$ blocks.

(c) Sort merge join $R \bowtie S$ using $M = 100$ blocks, assuming that if the second stage of the sort requires less than M blocks, then it can be combined with the merge-join step.

(d) Block-nested loop join $R \bowtie S$ using $M = 101$ blocks.

(e) Hash based join $R \bowtie S$ using $M = 100$ blocks. Assume both R and S are hashed using a hash function that distributes the tuples to hash buckets uniformly. Each hash bucket is joined using block-nested loop join.

Question 2. What is the cost of following index scan operations given the information below:

```
PAGES(S)=100,000  TUPLES(S)=4,000,000
Index I1 on S(B) with 5,000 leaf nodes and 3 levels (root/internal/leaf)
Index I2 on S(B,C) with 20,000 leaf nodes and 3 levels (root/internal/leaf)
Index I3 on S(C) with 5,000 leaf nodes and 3 levels (root/internal/leaf)
Index I4 on S(C,A,B) with 30,000 leaf nodes and 3 levels (root/internal/leaf)

Q1: SELECT A,B,C FROM S WHERE B>21 AND C=12180;
Q2: SELECT B FROM S WHERE C=12180;
Q3: SELECT A,B FROM S WHERE A LIKE 'A%' AND B>21 AND C=12180;
Q4: SELECT D FROM S WHERE A LIKE 'A%' AND C=12180;
Q5: SELECT A FROM S WHERE B>21;
```

Condition	Number of tuples expected
A LIKE 'A%'	8,000
B>21	200,000
C=12180	80,000
A LIKE 'A%' AND B>21	400
B>21 AND C=12180	4,000
A LIKE 'A%' AND C=12180	160
A LIKE 'A%' AND B>21 AND C=12180	8

Fill out the following table to estimate the cost of different query plans. If a certain index does not match a query (no conditions on the indexed attributes), simply write `N/A` in that cell.

The cost of an index scan includes the cost of scanning the index (internal and leaf nodes) and the cost of reading the matching tuples from disk. For the matching tuples, assume each tuple is in a different disk page in the worst case.

	Q1	Q2	Q3	Q4	Q5
Cost of sequential scan					
Condition for scanning I1					
Cost of index scan using I1					
Condition for scanning I2					
Cost of index scan using I2					
Condition for scanning I3					
Cost of index scan using I3					
Condition for scanning I4					
Cost of index scan using I4					

Question 3. What is the estimated number of tuples expected in the output of the following queries given the following statistics (note that VALUES is DISTINCT values for that attribute):
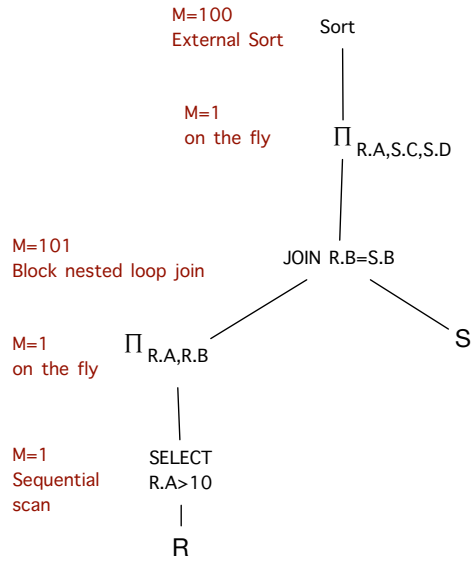
```
TUPLES(T)=200,000      PAGES(T) = 10,000
TUPLES(V)=10,000,000   PAGES(V) = 20,000

Q1: SELECT * FROM T WHERE T.X=10 ;
Q2: SELECT * FROM T WHERE T.Y>100 AND T.Y<200 ;
Q3: SELECT * FROM T WHERE T.X<>10 OR T.Y=20;
Q4: SELECT * FROM V WHERE V.X=10 ;
Q5: SELECT * FROM V,T WHERE V.X=T.X ;
Q6: SELECT * FROM V,T WHERE V.Y=T.Y ;
Q7: SELECT * FROM V,T WHERE V.Y=T.Y AND T.Z='ABC123' ;
Q8: SELECT * FROM V,T WHERE V.Y=T.Y AND V.X=T.X ;
```

Attribute	VALUES	Minval	Maxval
T.X	20,000	1	20,000
T.Y	2,000	1	4,000
T.Z	5,000	'AAA111'	'ZZZ999'
V.X	15,000	1	20,000
V.Y	3,000	1	4,000

Question 4. What is the total cost of the following query plan? The size of each relation after a specific operation is given below. Show the cost of each operation and explain with a single sentence.

Relation	Size in Pages
R	10,000
$\sigma_{R.A>10}(R)$	2,000
$\Pi_{R.A,R.B}(\sigma_{R.A>10}(R))$	80
S	20,000
$(\Pi_{R.A,R.B}(\sigma_{R.A>10}(R))) \bowtie_{R.B=S.B} S$	5,000
$\Pi_{R.A,S.C,S.D}(\Pi_{R.A,R.B}(\sigma_{R.A>10}(R))) \bowtie_{R.B=S.B} S$	400

M=100
External Sort   Sort

M=1
on the fly Π R.A,S.C,S.D

M=101
Block nested loop join   JOIN R.B=S.B

M=1 Π R.A,R.B   S
on the fly

M=1   SELECT
Sequential   R.A>10
scan

R

Question 5. Suppose you are given the following schedules. List all the conflicts and check whether they are serializable or not. Show your work.

(a) $S_1 : r_1(x)\ r_2(z)\ r_1(y)\ w_2(w)\ w_2(z)\ r_3(z)\ r_3(x)\ w_1(x)\ w_3(x)\ r_1(w)\ w_1(y)\ w_3(z)$

(b) $S_2 : r_1(x)\ r_1(y)\ r_2(z)\ r_3(z)\ r_3(x)\ w_1(x)\ w_3(x)\ w_2(w)\ w_2(z)\ r_1(w)\ w_1(y)\ w_3(z)$

Question 6. You are given the following schedule:

$$S : r_1(x) \; r_1(y) \; r_2(y) \; r_3(x) \; r_3(z) \; w_3(z) \; commit_3 \; w_1(x) \; commit_1 \; w_2(y) \; commit_2$$

Is this schedule possible using Strict Two Phase Locking with shared and exclusive locks? If yes, show which transactions obtain which locks when. If no, explain why a necessary locks cannot be obtained for this schedule to occur.