Database Systems, CSCI 4380-01
Homework # 8
Due Monday December 3, 2018 at 11:59 PM

Introduction.

This homework is worth 5.5% of your total grade. If you choose to skip it, the final exam will be worth 5.5% more. If you complete all the homeworks after Exam #2, there will be a special bonus in terms of your final grade computation. Additionally, doing this work now is essential to doing well in the final exam.

This homework requires no programming, only pen/pencil on paper computations. So you will submit your answers as a PDF file on GRADESCOPE.

Question 1. Estimate the cost of following operations with the information given below. All costs should be in number of pages (read or written).

Recall that when the final result is found, it is put in the output buffer. This operation has no additional cost.

```
R(A,B,C,D,E)  TUPLES(R)= 100,000   PAGES(R)= 2,000
S(F,G,A)      TUPLES(S) = 3,000,000 PAGES(S)= 6,000
```

(a) Block-nested loop join for `R ⋈ S` with `M=101` blocks. For join ordering, choose the lowest cost one and only show the result of that join.

(b) External sorting of `S` using `M=60` blocks.

(c) External sorting of `S` using `M=100` blocks.

(d) Hash join for `R ⋈ S` with `M=101` blocks. Describe how join works and any assumptions you make in your computation.

Assume tuples in R and S will uniformly distribute to 101 blocks. If a buckets after hashing will not fit in memory, then you will use block nested loop join.

(e) Sort merge join for `R ⋈ S` with `M=101` blocks. Sort each relation first and then join.

Note that if the number of partially sorted groups is smaller than allocated memory (in this case 101 blocks), they can be sorted and joined together in a single step.

Question 2. Estimate the cost of query `Q1` using different query plans. All costs should be in number of pages. Assume sufficient amount of memory is allocated for each query plan.

```
R(A,B,C,D,E)  TUPLES(R)= 100,000   PAGES(R)= 2,000
Q1: SELECT A,B FROM R WHERE C>10 AND D=25

Index I1 with three levels (root,internal,leaf) on R(C,D,A) with 800
                nodes in the leaf level

Index I2 with three levels (root,internal,leaf) on R(D,A,B,C) with 1,500
                nodes in the leaf level

Index I3 with three levels (root,internal,leaf) on R(C) with 300
                nodes in the leaf level

Index I4 with three levels (root,internal,leaf) on R(D) with 250
                nodes in the leaf level


Expected number of tuples of R for conditions:

Tuples(R.C>10)=20,000
Tuples(R.D=25)=1,000
Tuples(R.C>10 and R.D=25)=50
```

(a) Plan 1: sequential scan over `R`.

(b) Plan 2: using index `I1`.

(c) Plan 3: using index `I2`.

(d) Plan 4: using index `I3`.

(e) Plan 5: using index `I4`.

(f) Plan 6: using index `I3` and `I4` both.

Question 3. Estimate the size (number of tuples) of the following queries:

```
Q1:  select * from games where id = 21;
Q2:  select * from contestants where gameid = 2345;
Q3:  select * from contestants where shortname = 'Gilbert';
Q4:  select * from contestants where gameid = 2345 and shortname = 'Gilbert';
Q5:  select * from games g, contestants c
     where g.gameid=c.gameid and c.shortname='Gilbert';
Q6:  select * from responses where iscorrect = True;
Q7:  select * from responses where shortname = 'Gilbert' and gameid='5881';
Q8:  select * from responses where shortname = 'Gilbert' or gameid='5881';
Q9:  select * from responses r, contestants c
     where c.shortname=r.shortname;
Q10: select * from responses r, contestants c
     where c.shortname=r.shortname and c.gameid=r.gameid;
```

using the following statistics (VALUES is the same DISTINCT values):

```
  Tuples(Games) = 10,000
  Values(Games.gameid) = 10,000
  Values(Games.id) = 40
  Values(Games.airdate) = 10,000

  Tuples(Contestants)=30,000
  Values(Contestants.shortname)=3,000
  Values(Contestants.fullname)=18,000
  Values(Contestants.gameid)=10,000

  Tuples(Responses)=740,000
  Values(Responses.gameid)=10,000
  Values(Responses.clueid)=520,000
  Values(Responses.shortname)=3,000
  Values(Responses.iscorrect)=2
```

These are close to the real statistics, but they are not identical. However, the following is a great educational exercise (not required for solving the homework). You can get the statistics for the real data and compare the estimates to the real results (just by running select count(*) queries). When do the estimates are far off from reality and why?

Question 4. Estimate the cost of the following query plans for the same query. Which one is the cheapest plan?

The abbreviations used:

BNLJ	Block nested loop join
SMJ	Sort merge join
O-T-F	On the fly (pipelined operation)
Sort	External sort
I-O-S	Index only scan
SS	Sequential scan

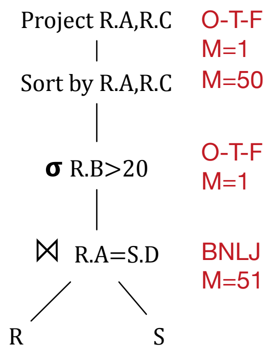
For simplicity, we will give you the size of the result of each operation. You can assume that there are appropriate projections to reduce the size of intermediate relations which are included in the given sizes.

Furthermore, assume that the join operation is over a foreign key (so for each S.D, there is a single R.A value). This simply makes the computation of sort merge join much simpler after a sort.

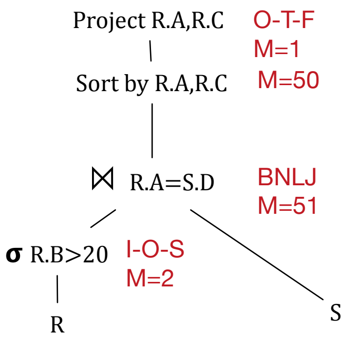
PAGES(R) = 100
PAGES(S) = 800
PAGES($\sigma_{R.B>20}$ R) = 25
PAGES(R $\bowtie_{R.A=S.D}$ S) = 700
PAGES(($\sigma_{R.B>20}$ R) $\bowtie_{R.A=S.D}$ S) = 175

Index scan for Plan 2 uses an index on `R.B,R.A,R.C`. Assume that index has 2 levels, root and 100 leaf pages. The selectivity of the condition $\sigma_{R.B>20}$ R is 1/4.

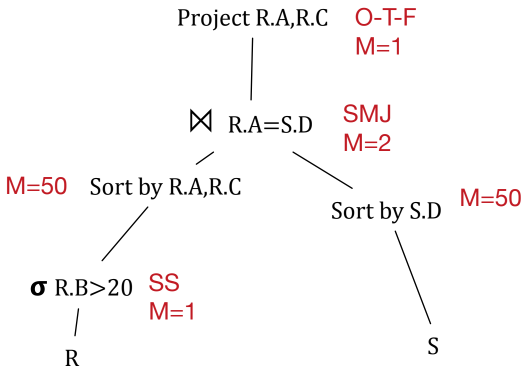
Note that operations in each query plan are pipelined from lower operations to upper operations, by placing the output of one operation to the input buffer of the next operation.



PLAN 1



PLAN 2



PLAN 3