CS411: Database Systems **Spring 2017**

Assignment 5

Due: Monday, April 24 at 11:59pm

General Instructions

- Feel free to talk to other members of the class in doing the homework. You should, however, write down your solutions yourself. *List the names of everyone you worked with at the top of your submission.*
- Keep your solutions brief and clear.
- Please use Piazza if you have questions about the homework but do not post answers. Feel free to use private posts or come to the office hours.

Homework Submission

- We DO NOT accept late homework submissions.
- We will be using Compass for collecting the homework assignments. Please submit your answers via Compass. Hard copies are not accepted.
- Contact the TAs if you are having technical difficulties in submitting the assignment; attempt to submit well in advance of the due date/time.
- The homework must be submitted in **pdf** format. Scanned handwritten and/or hand-drawn pictures in your documents won't be accepted.
- Please do not zip the answer document (PDF) so that the graders can read it directly on Compass. You need to submit one answer document, named as **hw5_netid.pdf**.
- Please see the assignments page for more details. In particular, we will be announcing errata, if any, on this page.

Question 1. Query Execution (10 points)

Consider the following relations:

- Relation R has 5,000 tuples, 250 tuples per block
- Relation S has 3,500 tuples, unknown number of tuples per block

The number of blocks in memory is 41 and the cost of joining R and S using a nested-loop join is 200.
Answer the following questions:

1. How many tuples per block does S have? (Do not forget to show your calculations.

   Answer1: R has 5,000/250 = 20 blocks (i.e. B(R) = 20) The cost for this join strategy is given by:

   Cost = B(R) + B(R) × B(S)/(M-2) (because B(S) >B(R)).

   Solving for B(S): B(S) = (Cost-B(R)) × (M-2) / B(R) = 162 blocks

   So S has 3600 tuples / 162 blocks = 22 tuples per block

   Answer2: Alternatively, Cost = B(R) + B(S) × ceil(B(R)/(M-2))

   solving for B(S) = 90

   Number of tuples per block = 3600/90 = 39

2. Using your answer above, what is the cost of joining R and S using the sort-merge algorithm?

   Answer1: 3*(B(R) + B(S)) = 3*(20 + 162) = 546

   Answer2: 3*(B(R) + B(S)) = 3*(20 + 39) = 330

3. What is the cost of joining R and S using a hash-based join?

   Answer1: 3*(B(R) + B(S)) = 3*(20 + 162) = 546

   Answer2: 3*(B(R) + B(S)) = 3*(20 + 39) = 330

4. Based on questions 2 and 3, explain which variant of the algorithm you would choose.

   Answer: Equivalently efficient

Question 2. Query Optimization (30 points)

Consider the relations A(x,y,z), B(w,x), and C(u,v,w), with the following properties: where, T(R) = number of tuples in relation R and V(R, a) = number of distinct values of attribute a in relation R. Estimate the sizes (measured in number of tuples) of the result of the following expressions:

1. A × C

   Answer: T(A) x T(C) = 4000 × 3000 = 12,000,000

A(x,y,z)	B(w,x)	C(u,v,w)
T(A) = 4000	T(B) = 1000	T(C) = 3000
V(A, x) = 30 V(A, y) = 30 V(A, z) = 40	V(B, w) = 250 V(B, x) = 50	V(C, u) = 10 V(C, v) = 40 V(C, w) = 100

2. A $\bowtie$ B

   Answer: T(A) x T(B) = 4000 $\times$ 1000 / max(V(A,x), V(B,x)) = 4000000/50 = 80000

3. `SELECT u FROM C WHERE u=20`

   Answer: T(C)/V(C,u) = 3000/10 = 300

4. $\sigma_{\text{x = 20 and u=30}}$ (B $\bowtie$ C)

   Answer: (T(B) $\times$ T(C)) / V(B $\bowtie$ C, x) V (B $\bowtie$ C, u) = 24

Question 3. Dynamic Programming (40 points)

Consider the following relations:

A(x,y,z)	B(w,x)	C(u,v,w)	D(u,z)
T(A) = 2500	T(B) = 1000	T(C) = 6000	T(D) = 2000
V(A, x) = 30	V(B, w) = 250	V(C, u) = 10	V(D, u) = 100
V(A, y) = 200	V(B, x) = 50	V(C, v) = 40	V(D,z) = 50
V(A, z) = 40		V(C, w) = 100	

We want to join all these relations as efficiently as possible. Determine the most efficient way to do the join. Clearly state any assumptions you have made. Show your work by completing the following table (each step in the dynamic programming algorithm should be one row):

Subset	Size	Lowest Cost	Lowest cost plan
...	...	...	...

Subset	Size	Lowest Cost	Lowest cost plan
AB	50,000		AB
BC	24,000		BC
CD	120,000		CD
AD	100,000		DA
ABC	1,200,000	24000	A(BC)
BCD	480000	24,000	(BC)D
ACD	6,000,000	100,000	(AD)C
ABD	2,000,000	50,000	(AB)D
ABCD	480,000	124,000	(BC)(AD)