Name: _____ **Student ID:** _____

FINAL EXAM

Fall 2022

CS 564 Introduction to Database Management Systems
Department of Computer Sciences
University of Wisconsin, Madison

Exam Rules:

1) Close book and notes, cheat sheet allowed, 90 minutes

2) Please write down your name and student ID number NOW.

3) Please wait until being told to start reading and working on the exam.

4) No electric devices are allowed.

1. (10 points) RELATIONAL ALGEBRA

(a) (5 points) List the names of the five basic operations of relational algebra.

(b) (5 points) Given relations Bars(bname, address, owner-name) and Sells(bar, beer, price), which can be joined on bname = bar, write a relational algebra expression that finds the addresses of all bars not owned by Mike that sell Bud for less than $3.

2. (20 points) SORTING

(a) (5 points) Why is a memory-based sort algorithm such as quick sort not appropriate for sorting a large amount of disk-based data?

(b) (15 points) Assume that I want to sort 24 data pages (on disk) using 4 buffer pages. How many passes do I need and what is the total cost? **Show your answer, that is, describe what happens in each pass, and how you derive the total cost. As usual, you should ignore the cost of writing out the final output.**

3. (15 points) RELATIONAL OPERATORS

In the following, M and N refer to the sizes (in pages) of the two relations that we want to join, and B refer to the size (in pages) of the memory buffer.

(a) (5 points) In the sort-merge join algorithm, what is the highest possible cost for the merging phase? Briefly describe a scenario when this cost occurs.

(b) (10 points) The typical cost of a sort-merge join algorithm is 5M + 5N. Here, the cost of sorting both relations is 4M + 4N, and the cost of merging the sorted relations is M + N. When we say that the cost of sorting both relations is 4M + 4N, we are making an assumption. State that assumption and explain why that leads to the cost 4M + 4N.

4. (25 points) QUERY OPTIMIZATION

(a) (5 points) In query optimization, is the goal is to find the plan that executes **the fastest**, among all possible plans?

(b) (5 points) For query optimization, why do we consider only left-deep join plans?

(c) (15 points) Consider two alternative plans to join three tables A, B, and C. The first plan is a left-deep join plan that first joins A with B (where A is the outer table), then join the output (of A joining B) with C, where C is the inner table.

The second plan is a right-deep join plan that first joins A with B (where A is the outer table), then join C with the output (of A joining B), where C is the outer table.

Assume that the joins are block nested loop joins. Assume each of the tables A, B, and C has 10 pages. Assume the output of joining A and B has 20 pages.

Assume that the memory M has 10 pages. Assume that when you do the left-deep join plan you can use 5 pages for the first join and 5 pages for the second join. Compute the cost of the above two plans. Explain how you arrive at these costs.

5. (10 points) CONCURRENCY CONTROL
   (a) (5 points) Briefly describe ACID properties.

(b) (5 points) Who is responsible for ensuring that the database remain consistent over time? The database system or the user who writes the application that uses the database?

6. (15 points) RECOVERY

Assuming we maintain a redo log whose content, after the crash, is:

<start T1>
<T1,X,3>
<start T2>
<T2,Y,1>
<start T3>
<T3,Z,5>
<T2,Y,2>
<commit T2>
<start ckpt(T1,T3)>
<T1,X,4>
<start T4>
<end ckpt>
<commit T1>
<T3,Y,2>
<T4,U,8>
<start ckpt(T3,T4)>
<start T5>
<T4,U,9>
<T5,X,10>
<commit T4>
<start T6>
<T6,Y,12>

Recover the database, assuming that after the crash the five elements have the values: X=1, Y=2, Z=3, U=4, V=5.

That is, list the values for X, Y, Z, U, and V after the recovery. Indicate which portion of the log you needed to inspect, and which transactions must be executed (that is, redone) again.

7. (5 points) NORMALIZATION
Define the notion of a BCNF table, that is, when do we say that a table is in BCNF?