CS4432: Database Systems II - Homework 3

April 24, 2025

Total Points: 100

Release Date: 04/24/2025 (11:00 AM)
Due Date: 05/05/2025 (11:59 PM)

Submission Instruction

1. Convert your file(s) to .pdf and upload it.
2. Include your name inside the file.
3. Submit your work on Canvas.
4. Late Submission Policy: Follows the policy posted on the course website.

Generative AI Usage Instruction

You are permitted to use generative artificial intelligence tools (such as ChatGPT, etc.) to assist in completing the assignment. However, when using these tools, you must clearly state the following in your assignment:

1. How the generative artificial intelligence tool was used.
2. How the generated answers were verified to be correct.

1 Problem 1 (Evaluation of Relational Operators) [70 Points]

Given the following relational operators and some properties about their input relations:

(a) Duplicate eliminator over unsorted relation R

(b) Grouping operator (group by column X) over a sorted relation R on column X

(c) Grouping operator (group by column X) over unsorted relation R

(d) Sorting operator (sort by column X) over unsorted relation R

(e) Sorting operator (sort by column X), and assume the operator can use a B-tree index that exists on R.X to read the tuples.

(f) Join of two relations R and S

(g) Bag Union of relations R and S

1) **[5 Points each Item]** For each of the items above, report whether the operator is **"Blocking"** or **"Non-Blocking"** and describe why.
//Remember that "Blocking" means the system cannot produce ANY output until it sees all the input.

2) **[5 Points each Item]** Assume relation R is 1,000 blocks and relation S is 150 blocks, and the available memory buffers are 200. Moreover, for Point (e) above, the R.X index size is 70 blocks. For each of the items above (a to g), discuss:

   a. Whether the operator can be done in one pass or not.

   b. If it can be done in one pass, what are the size constraints?

   c. If it cannot be done in one pass, then how many passes are needed? Describe the algorithm that uses the number of passes you suggest? What will be the I/O cost?

2 Problem 2 (Estimation of Relation Size) [30 Points (5 each)]

Given the following three relations $R1(a, b)$, $R2(b, c)$, and $R3(c, d)$ and associated statistics shown below in the meta data table. Estimate the number of tuples in the result relation for the different queries listed below, namely, $T(Q)$.

$$\begin{array}{lll} T(R1) = 400; & V(R1, a) = 50; & V(R1, b) = 50 \\ T(R2) = 500; & V(R2, b) = 40; & V(R2, c) = 100 \\ T(R3) = 1000; & V(R3, c) = 50; & V(R3, d) = 100 \end{array}$$

If there are any additional assumptions you need to make to answer any of the questions below, please explicitly state them.

1. $Q = \sigma_{(a=10)}(R1)$.
2. $Q = \sigma_{(a>=10)}(R1)$. (Assume that the range of $R1.a$ is $[1, 50]$).
3. $Q = \sigma_{(a>=10 \text{ AND } b=20)}(R1)$. Again assume the range of $R1.a$ is $[1, 50]$.
4. $Q = R1 \bowtie R2$, where $\bowtie$ represents natural join.
5. $Q = (R1 \bowtie R2) \bowtie R3$.
6. $Q = (\sigma_{(a>=10)}(R1)) \bowtie R2 \bowtie R3$.