

CS4432: Database Systems II

Spring D-Term

Homework 3

Release Date: April 15, 2014

Due Date: April 21, 2014 (11:59PM)

Total Points: 120

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

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

- (a) Duplicate elimination operator 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.
- 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), the R.X index size is 70 blocks. For each of the items above, 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 that number of passes you suggest? What will be the I/O cost?

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, $T(Q)$.

$$T(R1) = 400; V(R1, a) = 50; V(R1, b) = 50$$

$$T(R2) = 500; V(R2, b) = 40; V(R2, c) = 100$$

$$T(R3) = 1000; V(R3, c) = 50; V(R3, d) = 100$$

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. $Q = \sigma_{(a \geq 10)}(R1)$ (Assume that the range of $R1.a$ is $[1, 50]$).
3. $Q = \sigma_{(a \geq 10 \text{ AND } b=20)}(R1)$. 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. $Q = ((\sigma_{(a \geq 10)}(R1)) \bowtie R2) \bowtie R3$

Problem 3 (Query Processing Strategies & Their Costs) [20 Points (5 each)]

Consider the condition join $R1 \bowtie_{R1.a=R2.b} R2$, given the following information about the relations to be joined. The cost-metric is the number of IOs. The cost of writing the result would be the same independent of the particular join method used, hence we henceforth can ignore it. Given:

- R1 has 10,000 tuple, 10 tuples per block
- R2 has 2,000 tuple, 10 tuples per block
- The available memory buffers are 52

1. Assume we use a block-oriented nested loop join.
 - a. Which relation you suggest to be the outer relation?
 - b. What is the cost of the join if we use the outer relation as you suggested?
 - c. What is the cost of the join if we use the other relation (not what you suggestion) as the outer one?
2. Assume we use a sort-merge join, and we use the “Efficient Sort-Merge” algorithm covered in class where we merge the sorting and joining together
 - a. What is the cost of the join algorithm?
 - b. What is the minimum number of buffers needed for the cost to remain unchanged, i.e., Can we use less than 52 buffers and still have the same cost that you calculated in 2.a?
3. Assume we use a hash-join, and we will do a simple hash-join (not the complex hybrid-hash join).
 - a. What is the cost of the join algorithm?
 - b. What is the minimum number of buffers needed for the cost of the hash join to remain unchanged, i.e., Can we use less than 52 buffers and still have the same cost that you calculated in 3.a?
4. Assume we use an index-join with R2 as the outer relation, and we have an index on R1.
 - a. Assume that the index fits in memory. Moreover, on average we get 5 R1 tuples matching every R2 tuple.
 - a. What is the cost of the join algorithm?

What to Submit

- Include your answers in one file (.doc, .docx, or .pdf). This is the only file to submit.
- Include your name inside the file.

Where to Submit

- In WPI blackboard system

Late Submission Policy

- Follows the policy posted on the course website.