Name

CWID

Quiz
2

Due Saturday April 25th, 11:59am

CS525 - Advanced Database Organization
Spring 2020

Grading Guidelines

Please leave this empty! 1 2 Sum

Instructions

- **You have to upload the assignment to the course blackboard**.
- This is an individual and not a group assignment. Fraud will result in **0** points
- For your convenience the number of points for each part and questions are shown in parenthesis.
- There are **2** parts in this quiz
  1. Result Size Estimations
  2. I/O Cost Estimation

Part 1 Result Size Estimations (Total: 25 Points)

Consider a table `student` with attributes `CWID`, `name`, `major`, `credits`, a table `course` with `title`, `instructor`, `credits`, and a table `registered` with attributes `student` and `course`. `registered.student` is a foreign key to `CWID`. Attribute `course` of relation `registered` is a foreign key to attribute `title` of relation `course`. Given are the following statistics:

T(student) = 40,000	T(course) = 1,000	T(registered) = 20,000
V(student,CWID) = 40,000	V(course,title) = 1,000	V(registered, student) = 5,000
V(student,name) = 39,600	V(course,instructor) = 500	V(registered, course) = 100
V(student,major) = 50	V(course,credits) = 59	
V(student,credits) = 59		

Assume the range [0,90] for the `credits`.

Question 1.1 Estimate Result Size (5 Points)

Estimate the number of result tuples for the query $q = \sigma_{instructor=Bob}(course)$ using the first assumption presented in class (values used in queries are uniformly distributed within the active domain).

Solution

$$T(q) = \frac{T(course)}{V(course, instructor)} = \frac{1,000}{500} = 2$$

Correction Guideline

5 Points for showing or using the correct formula

Question 1.2 Estimate Result Size (5 Points)

Estimate the number of result tuples for the query $q = \sigma_{major=CS \vee major=EE}(students)$ using the first assumption presented in class.

Solution

$$T(q) = \frac{2 \cdot T(student)}{V(student, major)} = \frac{2 \cdot 40,000}{50} = 1,600$$

This answer is acceptable too. In case, students assume both predicates are independent:

$$T(q_1) = (1 - [(1 - \frac{1}{V(student, major)}) \cdot (1 - \frac{1}{V(student, major)})]) \cdot T(student)$$

$$= (1 - (1 - \frac{1}{50}) \cdot (1 - \frac{1}{50})) \cdot 40,000 = 1,584$$

Correction Guideline

5 Points for showing or using the correct formula

Question 1.3 Estimate Result Size (7 Points)

Estimate the number of result tuples for the query $q = \sigma_{credits \geq 39 \wedge credits \leq 44}(course)$ using the first assumption presented in class.

Solution

$$T(q) = \frac{(44 - 39 + 1) \cdot T(course)}{max(course, credits) - min(course, credits) + 1} = \frac{6 \cdot 1000}{90 - 0 + 1} = 66$$

Correction Guideline

7 Points for correct formula

Question 1.4 Estimate Result Size (8 Points)

Estimate the number of result tuples for the query q below using the first assumption presented in class.

$$q = \sigma_{credits \geq 30}(student) \bowtie_{CWID=student} registered \bowtie_{course=title} course$$

Solution

$$q_1 = \sigma_{credits \geq 30}(student)$$

$$T(q_1) = \frac{max(student, credits) - 30 + 1}{max(student, credits) - min(student, credits) + 1} \cdot T(student)$$

$$= \frac{90 - 30 + 1}{90 + 1} \cdot 40,000 = 26,813$$

$$q_2 = q_1 \bowtie_{CWID=student} registered$$

$$T(q_2) = \frac{T(q_1) \cdot T(registered)}{max(V(q_1, CWID), V(registered, student))}$$

$$= \frac{26,813 \cdot 20,000}{max(26,813, 5,000)} = 20,000$$

Thus,

$$q = q_2 \bowtie_{course=title} course$$

$$T(q) = \frac{T(q_2) \cdot T(course)}{max(V(q_2, course), V(course, title))}$$

$$= \frac{20,000 \cdot 1,000}{max(100, 1,000)} = 20,000$$

Correction Guideline

8 Points for correct formula

Part 2 I/O Cost Estimations (Total: 30 Points)

Question 2.1 External Sorting (5 Points)

You have a block of size of `128KB` and `2GB` memory. Compute the maximum size of file that can be sorted by `3-Pass Multiway Sort`.

Solution

$$M = \frac{2,097,152k}{128k} = 16,384 \text{ buffers}$$

$$\begin{aligned} B(R) &\leq M \cdot (M-1)^2 \text{blocks} \\ &\leq 16,384 \cdot (16,384-1)^2 \\ &\leq 16,384 \cdot (16,384-1)^2 \cdot 128K \end{aligned}$$

Correction Guideline

5 Points for showing or using the correct formula

Question 2.2 I/O Cost Estimation (25 Points)

Consider two unordered, non-clustered relations `R` and `S` with `B(R) = 200,000` and `B(S) = 2,000` blocks, and `S(R) =` $\frac{1}{20}$, and `S(S) =` $\frac{1}{10}$. You have `M = 101` memory pages available. Let `R` has an index on the joining attribute `C`. Compute the minimum number of I/O operations needed to join these two relations using **tuple-based-nested-loop join** (relations are not clustered), **block-nested-loop join** (relations are clustered), **merge-join** (the inputs are not sorted but contiguous), and **non-clustering index-join** (read of `S` with uniform distribution assumption and expected `20` matching tuples on `R` for `S`) and **hash-join** (relations are contiguous but not sorted). You can assume that the hash function evenly distributes keys across buckets. Justify you result by showing the I/O cost estimation for each join method.

Solution

- $T(R) = 200,000 \times 20 = 4,000,000$
- $T(S) = 2,000 \times 10 = 20,000$
- **TNL**:
  - `S`: Smaller relation as outer relation (better performance)
    $T(S) \times \big(1 + T(R)\big) = 20,000 \times (1 + 4,000,000) = 80,000,020,000$ I/Os
  - or `R`: Larger relation as outer relation
    $T(R) \times \big(1 + T(S)\big) = 4,000,000 \times (1 + 20,000) = 80,004,000,000$ I/Os
- **BNL**: S is smaller, thus, keep chunks of S in memory.
  - 100 buffers for S, 1 buffer for R
  - Cost for each S chunk:
    read chunk: 100 IOs
    read R: 200,000 IOs
  - $\frac{B(S)}{M-1} = \frac{2,000}{100} = 20$ S chunks
  - Total I/O cost is $20 \times (100 + 200,000) = 4,002,000$ IOs
- **MJ**:
  - Sort cost R by 3-Pass Multiway Join algorithm $= 6 \times B(R)$
  - Sort cost S by 2-Pass Multiway Join algorithm $= 4 \times B(S)$
  - Total cost = sort cost+join cost $= 7 \times B(R) + 5 \times B(S) = 7 \times 200,000 + 4 \times 2,000 = 1,410,000$ I/Os
- **IJ**: $T(S) + T(S) \cdot \frac{T(R)}{V(R,C)} = 20,000 + (20,000 \cdot 20) = 420,000$ I/Os
- **HJ**: $3 \cdot (B(R) + B(S)) = 3 \cdot (200,000 + 2,000) = 606,000$ I/Os

Correction Guideline

5 Points for each subquestion.