INFO20003 Semester 1, 2018

Assignment 3 – Query Processing

Due: Friday 18th May 2018 11:59pm AEST

Submission: Via LMS https://lms.unimelb.edu.au

Weighting: 10% of your total assessment. The assignment will be graded out of 20 marks.

Question 1 (5 marks)

Consider two relations A and B. A has 10,000 tuples, and B has 200,000 tuples. Both relations store 100 tuples per a page.

Consider the following SQL statement:
*SELECT **
FROM A, B
WHERE A.a = B.a;

We wish to evaluate a join between A and B, with an equality condition on A.a = B.a. There are 52 buffer pages available in memory for this operation. Both relations are stored as (unsorted) heap files. Neither relation has any indexes built on it. Assume that Sort-Merge Join can be done in 2 passes.

Consider alternative join strategies described below and calculate the cost of each alternative. Evaluate the algorithms using the number of disk I/O's (i.e. pages) as the cost. For each strategy, provide the formulae you use to calculate your cost estimates.

a) Page-oriented Nested Loops Join. Consider A as the outer relation. **(1 mark)**
b) Block-oriented Nested Loops Join. Consider A as the outer relation. **(1 mark)**
c) Sort-Merge Join. **(1 mark)**
d) Hash Join. **(1 mark)**
e) What would be the lowest possible I/O cost for joining A and B using any join algorithm and how much buffer space would be needed to achieve this cost? Explain briefly. **(1 mark)**

Question 2 (5 marks)

Consider a relation with the following schema:
Managers (id: integer, name:string, title:string, level: integer)

The Managers relation consists of 500,000 tuples stored in disk pages. The relation is stored as a simple heap file and each page stores 100 tuples. Possible titles in the Managers hierarchy are “CFO”, “CEO”, “CTO”, “Architect”, and “Team Lead”. Manager levels are ranging from 0-20.
Suppose that the following SQL query is executed frequently using the given relation:

SELECT name
FROM Managers
WHERE title = “Architect” and level > 18;

Your job is to analyse the access paths given below and estimate the cost of the best access path utilizing the information given about different indexes in each part.

a) Compute the estimated result size and the reduction factors (selectivity) of this query. **(1 mark)**
b) Compute the estimated cost of the best access path assuming that a *clustered B+ tree* index on *(title, level)* is (the only index) available. Suppose there are 200 index pages. Discuss and calculate alternative access paths. **(1 mark)**
c) Compute the estimated cost of the best access path assuming that an *unclustered B+ tree* index on *(level)* is (the only index) available. Suppose there are 200 index pages. Discuss and calculate alternative access paths. **(1 mark)**
d) Compute the estimated cost of the best access path assuming that an *unclustered Hash* index on *(title)* is (the only index) available. Discuss and calculate alternative access paths. **(1 mark)**
e) Compute the estimated cost of the best access path assuming that an *unclustered Hash* index on *(level)* is (the only index) available. Discuss and calculate alternative access paths. **(1 mark)**

Question 3 (10 marks)

Consider the following relational schema and SQL query. The schema captures information about employees, departments, and company finances. The finance sector is organized on a per department basis, i.e. each department has its own financial budget (and thus the corresponding record in the Finance relation).

Emp(eid: integer, did: integer, sal: integer, hobby: char(20))
Dept(did: integer, dname: char(20), floor: integer, phone: char(10))
Finance(did: integer, budget: real, sales: real, expenses: real)

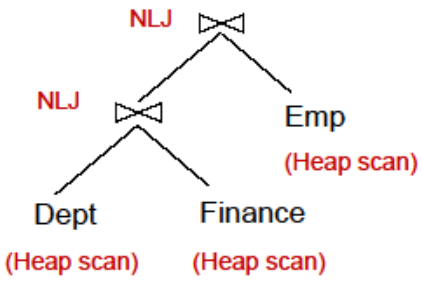
Consider the following query:

SELECT D.dname, F.budget
FROM Emp E, Dept D, Finance F
WHERE E.did=D.did AND D.did=F.did
AND E.sal ≥ 90000 AND E.hobby = 'bungee jumping;

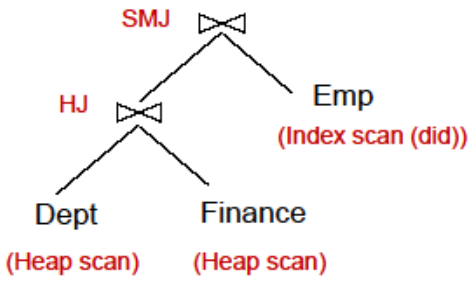
The system's statistics indicate that employee salaries range from 50,000 to 100,000, and employees enjoy 200 different hobbies. There are a total of 10,000 employees and 5,000 departments (each department with a corresponding financial record in the Finance relation) in the database. Each relation fits 100 tuples in a page. Suppose there exists a *clustered B+ tree* index on *(Emp.did)* of size 50 pages. Since selection over filtering predicates is not marked in the plans, assume it will happen later on-the-fly after all joins are performed.

a) Compute the estimated result size and the reduction factors (selectivity) of this query **(2 marks)**
b) Compute the cost of the plans shown below. Assume that sorting of any relation (if required) can be done in *2 passes.* NLJ is a *Page-oriented* Nested Loops Join. Assume that *did* is the candidate key, and that 100 tuples of a resulting join between Emp and Dept fit in a page. Similarly, 100 tuples of a resulting join between Finance and Dept fit in a page. **(8 marks, 2 marks per plan)**

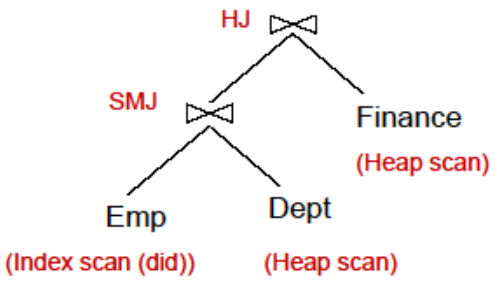
1)



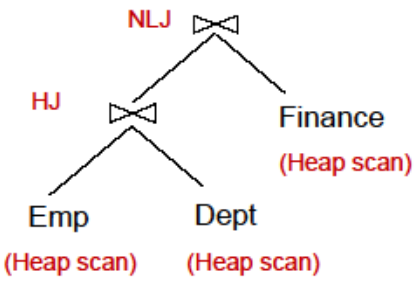
2)



3)



4)



Formatting Requirements:

For each question, present an answer in the following format:

- Show the question number and question in **black** text
- Show your answer in **blue** text (please type your answers on a computer)
- For each of the calculations provide the formulae you used to calculate your cost estimates (not only the result)

Submission Process:

Submit a single PDF showing your answers to all questions to the Assessment page on LMS by midnight on the due date of Friday 18th of May. Name your file 'STUDENT_ID'.pdf, where STUDENT_ID corresponds to YOUR student id.

Requesting a submission deadline extension:

If you need an extension due to a valid (medical) reason, you will need to provide evidence to support your request by *Thursday 17th of May 5pm.* Medical certificates need to be at least 2 days in length.

To request an extension:

1. Email the Subject Coordinator (deccles@unimelb.edu.au) with your student id, your name and your university email with the extension request and supporting evidence.
2. If your submission deadline extension is granted you will receive an email reply granting the new submission date. Do not lose this email!

Reminder: INFO20003 Hurdle Requirements

To pass INFO20003 you must pass two hurdles:

- Hurdle 1: Obtain at least 50% (15/30) or higher for the three individual assignments (each worth 10%)
- Hurdle 2: Obtain a grade of 50% (35/70) or higher for the Mid Semester Test (10) and the End of Semester Exam (60)

Therefore, it is our recommendation to students that you attempt every assignment and every question in the MST and exam.

Good Luck!