INFO20003 Semester 1, 2020

Assignment 3 – Query Processing and Query Optimisation
Date Due: Friday 5th June 2020 06:00pm AEST
Submission: Via LMS: https://canvas.lms.unimelb.edu.au
Weighting: 10% of your total assessment

Question 1 (5 marks)

Consider two relations `A` and `B`. `A` has 80,000 tuples, and `B` has 100,000 tuples. Both relations store 100 tuples per page. Consider the following SQL statement:

```
SELECT *
  FROM A INNER JOIN B
    ON A.a = B.a;
```

We wish to evaluate an equijoin between `A` and `B`, with an equality condition `A.a = B.a`. There are 102 buffer pages available for this operation. Both relations are stored as (unsorted) heap files. Neither relation has any indexes built on it.

Consider the 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. Assume that Sort-Merge Join can be done in 2 passes. (1 mark)
d) Hash Join (1 mark)
e) What would be the lowest possible cost to perform this query, assuming that no indexes are built on any of the two relations, and assuming that sufficient buffer space is available? What would be the minimum buffer size required to achieve this cost? Explain briefly. (1 mark)

Question 2 (5 marks)

Consider a relation with the following schema:

`JobSeekers(id, firstname, lastname, city, soughtsalary)`

The `JobSeekers` relation consists of 10,000 pages. Each page stores 100 tuples. The online software works in the 8 largest Australian cities and `soughtsalary` can have values between 60,000 and 160,000 (i.e., [60,000 – 160,000].)

Suppose that the following SQL query is executed frequently using the given relation:

```
SELECT *
  FROM JobSeekers
 WHERE city = 'Melbourne' AND soughtsalary > 80,000;
```

Your job is to:

a) Compute the reduction factors and the estimated result size in number of tuples. **(1 mark)**
b) Compute the estimated cost in number of disk I/O's of the best plan if a *clustered B+ tree* index on *(*city, soughtsalary*)* is the only index available. Suppose there are 2,000 index pages. Discuss and calculate alternative plans. **(1 mark)**
c) Compute the estimated cost in number of disk I/O's of the best plan if an *unclustered B+ tree* index on *(*soughtsalary*)* is the only index available. Suppose there are 2,000 index pages. Discuss and calculate alternative plans. **(1 mark)**
d) Compute the estimated cost in number of disk I/O's of the best plan if an *unclustered Hash* index on *(*city*)* is the only index available. Discuss and calculate alternative plans. **(1 mark)**
e) Compute the estimated cost in number of disk I/O's of the best plan if an *unclustered Hash* index on *(*soughtsalary*)* is the only index available. Discuss and calculate alternative plans. **(1 mark)**

Question 3 (10 marks)

Consider the following relational schema and SQL query. The schema captures information about employees, departments, and company finances (organized on a per department basis).

```
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:

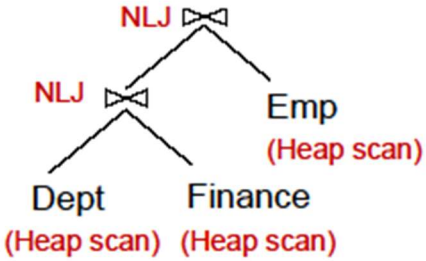
```
SELECT D.dname, F.budget
  FROM Emp E, Dept D, Finance F
 WHERE E.did = D.did
   AND D.did = F.did
   AND E.sal > 100,000
   AND E.hobby IN ('diving', 'soccer');
```

The system's statistics indicate that employee salaries range from 50,000 to 150,000, and employees enjoy 50 different hobbies. There is a total of 25,000 employees and 1,200 departments (each with 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 (Dept.did) and a clustered B+ tree index on (Emp.salary), both of size 50 pages.

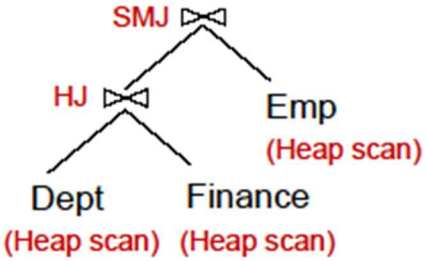
a) Compute the reduction factors and the estimated result size in number of tuples. **(2 marks)**

b) Compute the cost in number of disk I/O's 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 50 tuples of a resulting join between Emp and Dept fit in a page. Similarly, 50 tuples of a resulting join between Finance and Dept fit in a page. Any selections/projections not indicated on the plan are performed "on the fly" **after** all joins have been completed. **(8 marks, 2 marks per plan)**

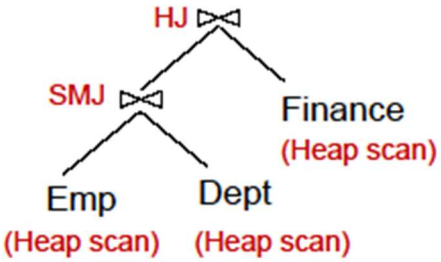
1)


NLJ
NLJ
Emp
(Heap scan)
Dept
(Heap scan)
Finance
(Heap scan)

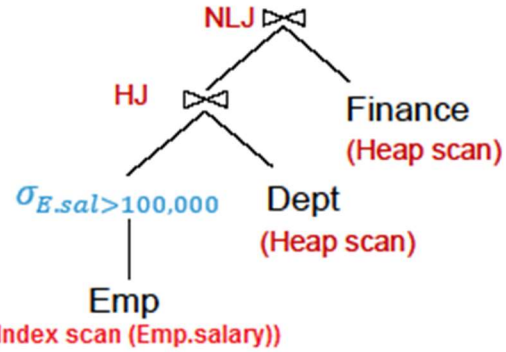
2)


SMJ
HJ
Emp
(Heap scan)
Dept
(Heap scan)
Finance
(Heap scan)

3)


HJ
SMJ
Finance
(Heap scan)
Emp
(Heap scan)
Dept
(Heap scan)

4)


NLJ
HJ
Finance
(Heap scan)
$\sigma_{E.sal>100,000}$
Dept
(Heap scan)
Emp
(Index scan (Emp.salary))

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
- Start Question 2 and Question 3 on a new page.

For each of the calculations provide the formulae you used to calculate your cost estimates

Submission Process:

Submit a single PDF showing your answers to all questions to the Assessment page on LMS by 6pm on the due date of Friday 5th of June. 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 9pm, Thursday 4th of June. Medical certificates need to be at least two days in length.

To request an extension:

1. Email Oscar Correa (oscar.correa@unimelb.edu.au) from your university email address, supplying your student ID, 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! Replies may take up to 12 hours, so please be patient.

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 assignments (each worth 10%)
• Hurdle 2: Obtain a grade of 50% (35/70) or higher for the End of Semester Exam

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

GOOD LUCK!