Data Management (A.A. 2023/24) – exam of 11/07/2024

Problem 1 We refer to a setting with both shared and exclusive locks. A transaction T is called *cautious* if all its lock requests appear before every other action of T. A schedule is called *cautious* if it is legal and all its transactions are both well-formed and cautious. A lock-based, active scheduler with both shared and exclusive locks is called *heedful* if it behaves like a 2PL scheduler with the only addition that it does not accept schedules that are not cautious (where we adopt the usual definition of an active 2PL scheduler "accepting" a schedule). Prove or disprove the following sentences.

1.1 Every cautious schedule is a 2PL schedule.

1.2 Every cautious schedule is conflict serializable.

1.3 Every heedful sheduler can avoid to deal with deadlock management.

Problem 2 Consider the following transactions: $\boldsymbol{T_1} : r_1(A)\, w_1(C)$; $\boldsymbol{T_2} : r_2(B)\, w_2(A)$; $\boldsymbol{T_3} : r_3(C)\, w_3(D)$; $\boldsymbol{T_4} : w_4(D)\, r_4(B)$ and answer the following questions, providing a detailed justification for each answer.

2.1 How many non-serializable schedules do exist on T_1, T_2, T_3, T_4?

2.2 Prove or disprove that all serializable schedules on T_1, T_2, T_3, T_4 are accepted by the 2PL scheduler with both shared and exclusive locks.

Problem 3 Given the table `R(A,B,C)` with 1000 tuples stored in a heap with 100 pages and given 80 frames in the buffer, we want to answer the query:

```
select A, count(*) from R group by A having count(*) > 10
```

3.1 Illustrate in detail a multi-pass algorithm to solve the problem, and tell which is its cost.

3.2 Tell whether the block-nested loop technique is suitable for answering the query.

3.2.1 If the answer is positive, then illustrate the algorithm and tell which is its cost.

3.2.2 If the answer is negative, consider the following question: can we build and use an index during the application of the block-nested loop algorithm in such a way that the modified block-nested loop algorithm is able to produce the result of the query? If the answer to the question is negative, then motivate the answer in detail; if the answer is positive, provide a detailed description of the new algorithm and tell which is its cost.

Problem 4 We refer to a setting where the data manager has 3 buffer frames available, we have a table `R(A,B,C,D)` with 4.000 tuples stored in a heap with 200 pages and we want to compute the relation obtained from `R` by eliminating duplicates.

4.1 Illustrate in detail a multipass algorithm for solving the problem in the above mentioned setting and tell which is its cost.

4.2 Making use of a hash-based index, design an index-based algorithm for solving the problem in the above mentioned setting, provide a detailed description of such algorithm and tell which is its cost.

4.3 Making use of a B^*-tree-based index, design an index-based algorithm for solving the problem in the above mentioned setting, provide a detailed description of such algorithm and tell which is its cost.

Problem 5 (only for students who do **not** do the project)
Consider the relations Drone(num,model,date) with 1.000 pages and 30.000 tuples, and Mission(code,num,model,date) with 4.000 pages and an associated index on Mission with search key ⟨model,date⟩, for which we know that the cost of retrieving the records with a specific value of attribute model is 6 page accesses. Assume a buffer with 100 frames, and consider the two queries shown below.

Query Q_1:

```
select num, model from Drone
except all  -- not removing duplicates
select num, model from Mission
```

Query Q_2:

```
select model, date from Drone
except all  -- not removing duplicates
select model, date from Mission
```

where "except all" denotes bag difference. For both queries Q_1 and Q_2, tell (i) whether it is possible to process the query by using a block-nested loop algorithm, and (ii) whether it is possible to process the query by an index-based algorithm using the above-mentioned index on Mission. In all four cases, if the answer is positive, then describe the algorithm and tell which is its cost. If the answer is negative, then motivate the answer in detail.