Data Management – AA 2017/18 – exam of 06/02/2018

Problem 1

Describe in the most possible detailed way the two-pass algorithm based on hashing for duplicate elimination, and indicate the condition under which such algorithm can be used, and its cost in terms of number of page accesses, providing an appropriate motivation for both answers.

Problem 2

Consider the relation SOCCERGAME(code,team1,team2,date,result), that for each soccer game stores information about its code, the teams participating in the game, the date and the result. The relation has 1.900.000 tuples, stored in 190.000 pages. We assume that we have 70 buffer frames available, that all fields and pointers have the same length, and that for every value of team1 there are 100 games in the relation. Also, there is a nonclustering B$^+$-tree index on SOCCERGAME with search key team1 using alternative 2. Consider the query

```
select s1.code, s2.code
from SOCCERGAME s1 join SOCCERGAME s2 on s1.team1 = s2.team2
where s1.team1 = 'Barcelona'
```

2.1 Describe the logical query plan associated to the query code, and illustrate both the logical and the physical query plan you would select, motivating the choices.

2.2 Tell which is the cost (in terms of number of page accesses) of executing the query according to the selected physical query plan.

Problem 3

Consider the relation FLIGHT(fcode,company,departure,destination), storing the code, the air company, and the cities of departure and destination of a set of flights, and the relation TAKES(flight,person,date,cost), recording the flights taken by the various persons in the various dates, with the corresponding cost. We know that TAKES has 2.000.000 tuples, each page contains 50 such tuples in the average, and there is a dense B$^+$-tree index on TAKES with search key ⟨flight,person⟩ using alternative 2. We also know that FLIGHT is stored in a file with 400.000 pages sorted on ⟨fcode,company⟩. Finally, we know that there are 100 buffer frames available, each flight is taken by 500 persons at most, and each value or pointer occupies the same space in memory. Consider the query

```
select person, company
from TAKES join FLIGHT on fcode = flight
where company = 'ALITALIA' or company = 'AIRFRANCE'
```

3.1 Describe the logical query plan associated to the query code, and illustrate both the logical and the physical query plan you would select, motivating the choices.

3.2 Tell which is the cost (in terms of number of page accesses) of executing the query according to the selected physical query plan.

Problem 4

Consider the following schedule

$$S = r_1(x)\, w_2(x)\, r_3(x)\, w_1(u)\, w_3(v)\, r_3(y)\, r_2(y)\, w_3(u)\, w_4(t)\, w_3(t)$$

4.1 Tell whether S is conflict-serializable. If the answer is no, then motivate the asnwer. If the answer is yes, then exhibit at least one serial sechelue conflict-equivalent to S.

4.2 Tell whether S is a 2PL schedule (with both shared and exclusive locks) or not, explaining the answer in detail.

4.3 Tell whether S follows the strict 2PL protocol (with shared and exclusive locks) or not, explaining the answer in detail.

Problem 5

Consider the three transactions T_1, T_2, T_3 defined as follows:

$\boxed{T_1 = r_1(A),\ w_1(A)}$ $\boxed{T_2 = r_2(A),\ w_2(A)}$ $\boxed{T_3 = r_3(A),\ w_3(A)}$

and answer the following questions, motivating the answer:

- How many non-serial schedules do exist on T_1, T_2 which are conflict serializable?
- Is there at least one non-serial schedule on T_1, T_2, T_3 that is view-serializable?
- Is there at least one non-serial schedule on T_1, T_2, T_3 that is strict?