FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO
Mestrado Integrado em Engenharia Informática e Computação

Database Technologies

OPTIMIZATION EXERCISES

Problem 1: Selection optimization

- Suppose that the relation R(A,B,C,D) has a clustering index on A and non-clustering indexes on the other columns. The four indexes have the images, resp., 50, 10, 20 e 100. The number of tuples in the relation is 10 000, occupying 500 blocks. Which are the ways of evaluating the query

$$\pi_A(\sigma_{A=0 \wedge B=1 \wedge C>2 \wedge D=3}(R))$$

and their cost? Which one is cheaper?

Problem 2: Selection optimization

Suppose that, with respect to a DB with a table PERSONS(Cod, Name, Category, Dept, Building, Office), you have the following data: the table has 10000 tuples, 125 bytes each, disk blocks are about 8000 bytes long, there is a clustering index on the key COD and non-clustering indexes on NAME and DEPT and the number of departments is 10. Estimate and justify the cost of evaluating

```
SELECT Cod, Name
FROM Persons
WHERE Cod <1000 AND Dept='DEEC'.
```

Problem 3: Data structures

Some data structures used in file processing provide clustering indexes and some provide non-clustering indexes. Among the data structures listed below which belong to the former, to the latter or to none of them, from the viewpoint of their ability to be used in the selection optimization algorithm studied?

a) Hash table with records stored in buckets.

b) Hash table with records stored in a heap, referenced by pointers in the buckets of the hash table.

c) B-tree with the records stored in the leaves.

d) B-tree with leaves pointing to linked lists of records with the same key.

Problem 4: Join optimization

Suppose you have two relations R(A,B), with 1 000 000 tuples and S(B,C) with 100 000. You can store in a block 20 R records or 100 S records and B image size is 500.

a) What is the output cost of computing $R \times S$?

b) What is the input cost of computing $R \times S$ if the available memory is M=100 blocks?

c) Repeat b) for the case M=1000. Comment the result.

d) What is the output cost for the natural join of R and S?

e) What is the cost of computing $R \bowtie S$ by the sort-join method, M=100 and M=1000?

f) What is the cost of computing $R \bowtie S$ based on a clustering index on S.B?

Problem 5: Join optimization

Consider the relations R(A,B,C), S(C,D,E) and T(E,F) with sizes R: 1000 tuples, S: 1500, and T:750.

a) Assume that the primary keys are respectively A, C and E. Estimate the size of the join $R \bowtie S \bowtie T$ and suggest an efficient strategy for its computation.

b) Assume that there are no primary keys (apart from the complete relation) and that the image sizes of the attributes are: R.C-900, S.C-1100, S.E-50, T.E-100. Estimate the size of the join $R \bowtie S \bowtie T$ and suggest an efficient strategy for its computation.

Problem 6: Join optimization

Suppose you have two relations R(A,B,C), with 1 000 000 tuples, and S(B,C,D), with 100 000 tuples. Assume that 20 R records or 100 S records fit in a disk block and that $R.B \subseteq S.B$ and $S.C \subseteq R.C$. The image sizes are R.B - 100, S.B - 200, R.C - 50 and S.C - 10.

a) Estimate the cost of computing the natural join by the sort-join method.

b) Would it be preferable to use a method based on two clustering indexes on R.B and S.B? How much?

Problem 7: Join optimization

Estimate the output size as a function of the sizes of the relations (T_R e T_S), supposing you want to perform the join of the relations R(A,B) and S(B,C), with the condition

a) $R.B < S.B$

b) $R.B <> S.B$

Problem 8: Algebraic manipulation

The equation $\pi_S(E_1-E_2) = \pi_S(E_1) - \pi_S(E_2)$ is valid? Justify with an example.