Advanced Database Management Systems

Date:  Tuesday 24th May 2011
Time:  14:00 - 16:00

Please answer BOTH questions

For full marks your answers should be concise as well as accurate.
In a question, the word BRIEFLY is used to indicate a number of sentences between 1 and 5, and never more than 10 or so.

Marks will be awarded for reasoning and method as well as being correct.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text.
1. a) Consider the following relation schemas:

```
suppliers (sid: integer, sname: string, address: string)
parts (pid: integer, pname: string, colour: string)
catalogue (sid: integer, pid: integer, cost: real)
```

The primary key of the relation comprises the underlined attributes, whilst the domain type of each attribute is stated after the colon. Thus, sid is the primary key for suppliers; pid, the primary key for parts; and the concatenation sid, pid, the composite primary key for catalogue, which states the price charged for a given part by a given supplier.

i) Write the following query in the extended relational algebra presented in this course unit: “Retrieve the minimum cost for each part, on a per colour basis”.

(3 marks)

ii) Briefly describe one possible optimization that you can apply (or have already applied) to the relational-algebraic expression you gave as an answer to the preceding subpart.

(3 marks)

iii) Draw a logical operator tree (or write the corresponding algebraic expression) that would classically result from the direct, clause-by-clause translation of the following SQL query into relational algebra:

```
select s.sname
from suppliers s, parts p, catalogue c
where s.sid = c.sid and p.pid = c.pid and c.cost > 100 and p.colour = "red"
```

(6 marks)

iv) Draw a logical relational operator tree (or write the corresponding algebraic expression) that shows the effect of pushing projections and selections as close to the leaves as possible on your answer to the preceding subpart.

(7 marks)

b) Recall the classical query optimization stages (i.e., parsing, translation, rewriting, and plan selection).

i) State in which of the stages above is the join ordering problem often addressed.

(2 marks)

ii) State whether (i.e., yes or no) there is potential benefit in addressing the join ordering problem in more than one stage. If your answer is yes, then indicate how the benefits provided in each stage should add up. Otherwise (i.e., if your answer is no), indicate how the benefits provided in each stage might not add up.

(5 marks)
c) Briefly contrast the impact of systematically long query optimization times (relative to total response time) in the case of classical queries and of continuous queries over streams.

(4 marks)
2. a) Consider again the following relation schemas (already shown and described above):

```
suppliers (sid: integer, sname: string, address: string)
parts (pid: integer, pname: string, colour: string)
catalogue (sid: integer, pid: integer, cost: real)
```

Now, assume that the estimated size of `suppliers` is a tenth of the estimated size of `parts` and a fifth of the estimated size of `catalogue`. Finally, assume that the database that stores them is distributed as indicated by the following join graph:

```
Site 1
Suppliers

sid

Site 2
Catalogue

pid

Site 3
Parts
```

i) Using the notation described in the course, state, step by step, the evaluation program that is heuristically most efficient, given the information above, for the following query injected at site 3:

```
select *
from suppliers s, parts p, catalogue c
where s.sid = c.sid and p.pid = c.cid
```

(5 marks)

ii) Briefly explain why an evaluation program that scheduled all the operators in the query for execution at site 3 is unlikely, heuristically speaking, to be more efficient under classical assumptions.

(2 marks)

b) Briefly explain why semijoins are often useful in distributed query evaluation and whether they have any downside.

(4 marks)

c) Consider the notion of data appliances introduced in this course unit.

i) Briefly explain what characteristics of this kind of data management technology the word `appliance` is aiming to highlight.

(2 marks)

ii) Assume that you are working as a member of a team that has been assigned the task of assessing the benefits of switching the analytical workload of the organization to a data appliance product. At one point in the discussions, Yu,
a member of the team, argues that the estimate for secondary memory needed is wrong because it disregards the fact that indices occupy a lot of space in disk. Rajiv, the team member who came up with the estimate replies to Yu as follows: “It's an appliance. This should tell you there's nothing wrong with the estimate.” At this point, the leader of the team asks you for your opinion. State whether you agree with Yu or Rajiv and briefly explain your reasons for doing so using the technical notions introduced in this course unit.

(4 marks)

d) Briefly relate the unbounded nature of data streams with the semantic characteristics of the selection and the projection operators in the relational algebra.

(4 marks)

e) Assume the following sensor network topology:

![Sensor Network Topology Diagram]

In the figure above, a vertex denotes a sensor node whose name is the vertex label (with B denoting the base station), an edge denotes that the nodes involved are within wireless communication range of one another (i.e., are one-hop neighbours), and the edge label denotes the quality of the communication link (the higher the label value, the better the quality).

Using the basic flooding algorithm discussed in this course unit, show the step-by-step derivation of a routing tree for the above network topology.

(9 marks)