Advanced Database Management Systems

Date: Thursday 28\textsuperscript{th} January 2010

Time: 14.00 – 16.00

Please answer BOTH questions

The use of electronic calculators is permitted provided they are not programmable and do not store text
1. 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 key fields are underlined, and the domain of each field is listed after the field name. Therefore sid is the key for Suppliers, pid is the key for Parts, and sid and pid together form the key for Catalogue. The Catalogue relation lists the prices charged for parts by Suppliers.

a) Write the following query in relational algebra: “Find the pids of green parts with Manchester suppliers.” (3 marks)

b) Very briefly explain in English what the following query in relational algebra computes:

\[ \Pi_{\text{address}}(\sigma_{\text{cost}>100}(\sigma_{\text{sname}=\text{ABC}}(\text{Suppliers}) \bowtie \text{Catalogue})) \]  (3 marks)

c) State two rewrite rules taught in this course unit that, heuristically speaking, could be used to derive a more efficient algebraic expression than the one in (1.b) above whilst briefly explaining what effect their application is intended to have. (6 marks)

Assume that a large retail company (that sells, say, clothes) has had a surge in sales. This was largely driven by online sales in the company’s e-commerce website and it has been claimed that the database backend is running the risk of becoming a bottleneck, with detrimental effects to the company’s business strategy. In particular, the company is worried about long-term solutions and with high maintenance costs. An external report was produced to discuss potential solutions that concluded by recommending the use of a Type 0 data appliance. The data management team opposed the report’s conclusion on the grounds that “The recommended strategy is wrong for our workloads.”. As a consultant, you were asked to arbitrate.

d) Briefly explain what is a data appliance and what being a Type 0 data appliance means. (2 marks)

e) State whether you agree or not with the data management team. (1 mark)

f) Briefly explain the reasons behind your opinion in (1.e) above. (3 marks)

In the context of peer-to-peer databases, use the operations associated with distributed hash tables to briefly explain how they implement:

g) An overlay network. (2 marks)

h) A storage framework. (2 marks)

(Question 1 continues on the following page)
Consider a database comprising the following relation schemas already shown above:

**Suppliers** (sid: integer, sname: string, address: string)
**Parts** (pid: integer, pname: string, colour: string)
**Catalogue** (sid: integer, pid: integer, cost: real)

Assume now that the database is distributed as indicated by the following join graph:

1. Using the notation described in the course unit, and on the basis of the information given immediately above, describe the evaluation program that you would recommend if the size relationships were *Catalogue* < *Parts* < *Suppliers*. (8 marks)
2. a) Briefly explain the contrasts between classical data management and stream data management suggested by the following issues:

i) How query evaluation differs (1 mark)
ii) How the blocking nature of a relational operator matters (2 marks)
iii) How important or not is the ordered nature or not of the inputs (3 marks)

b) Briefly explain what is meant by an asymmetric join strategy in the context of sliding window joins in stream query processing and cite one circumstance in which such a strategy may not prove useful. (2 marks)

c) Assume a sliding window equijoin is running over two windows X and Y (under bag semantics). Assume also that X is defined to hold the last 6 tuples to arrive in it and Y the last 4. Assume further that both tuples are unary with a single attribute of type string and are time-stamped (with integer values). Let the current state of the windows be as follows:

\[
X = [(a, 1), (b, 2), (d, 2), (a, 3), (e, 4), (a, 5)] \\
Y = [(b, 2), (c, 3), (a, 4), (a, 3)]
\]

Using the core semantics of a sliding window join, describe the steps that are carried out when:

i) (c, 5) arrives in X (when the state is as above). (3 marks)
ii) (a, 6) arrives in Y (when the state is as resulted from (i)). (3 marks)

Assume that, in a distributed database, a relation with the following schema

Projects (pno: integer, pname: string, budget: integer, loc: string)

is horizontally fragmented per project number pno according to the following rule Projects1 contains projects with numbers up to and including 1000; Projects2, those between and including 1001 and 6000; and Projects3, those above and including 6001.

d) Give the relational algebraic expression that defines each fragment. (3 marks)

e) Give a direct translation into relational algebra of the following SQL query over the distributed database:

\[
\begin{align*}
&\text{SELECT} & \text{AVG(P.budget)} \\
&\text{FROM} & \text{Projects P} \\
&\text{WHERE} & P.pno > 7000 \\
&\text{GROUP BY} & P.loc
\end{align*}
\]

(3 marks)
(Question 2 continues from the previous page)

f) Irrespective of other considerations that may bear on the decision, show that this query is amenable to being evaluated using a map-reduce engine (such as discussed in the course unit) by explaining how the computation carried out by the engine can be equated to the evaluation of some algebraic expressions.

(6 marks)

In the context of sensor network data management:

g) Briefly explain under which assumption(s) is sensor network query processing more efficient than the alternative of sending all the sensed data to be processed at the base station.

(2 marks)

h) Briefly explain in what way the use of aggregation trees is related to the underlying issue in (2.g) immediately above.

(2 marks)