Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Advanced Database Management Systems

Date: Thursday 16th January 2014
Time: 09:45 - 11:45

Please answer BOTH Questions

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) Briefly explain why the existence of equivalent calculi and algebra for querying a relational database is an important enabler of the approach to query optimization studied in this course unit (i.e., one centred on a stack of stages, each of which implements a specific transformation on a linguistic expression until an executable plan is obtained). (2 marks)

b) Given two relations $R_1(a, b, e)$ and $R_2(a, c, d)$, where $R_1.a$ is the primary key of $R_1$, $R_2.a$ is the primary key of $R_2$, $R_1$ contains $N_1$ tuples, $R_2$ contains $N_2$ tuples, and $N_2 > N_1 > 0$, explain what the minimum and maximum possible cardinalities (in terms of $N_1$, $N_2$, or both, as appropriate) for the resulting relation produced by each of the following relational-algebraic expressions might be. In each case, state any assumptions about the schemas of $R_1$ and $R_2$ that are needed to make the expression meaningful:

i) $R_1 \cap R_2$

ii) $\pi_{a,d}(R_2)$

(4 marks)

c) Consider the following relation schemas:

| suppliers (sid: int, sname: str, address: str) |
| parts (pid: int, pname: str, colour: str) |
| catalogue (sid: int, pid: int, cost: real) |

Assume that, in the database, the current extents of the relations above are as follows:

<table>
<thead>
<tr>
<th>suppliers</th>
<th>parts</th>
<th>catalogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>sid</td>
<td>sname</td>
<td>address</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>Y</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>Z</td>
</tr>
</tbody>
</table>

Assuming bag semantics and assuming that the join is the natural join, show (as tables, with the correct column names) each intermediate result, as well as the final result, of evaluating the following relational-algebraic expression as written:

$$\text{pid} \gamma_{\text{avg(cost)}} \rightarrow \text{avgcost} \left( \sigma_{\text{sid} = 5} \left( \pi_{\text{pid}, \text{sid}, \text{cost}} \left( \text{parts} \bowtie_{\text{pid}} \text{catalogue} \right) \right) \right)$$

(6 marks)
d) Consider again the suppliers, parts and catalogue relation schemas used in the preceding item. Write the following query in SQL: “Retrieve the address of suppliers that sell parts costing no more than £35.” (1 mark)

e) Again in the context of the suppliers, parts and catalogue relation schemas used in the preceding item, write the expression that results from the direct translation of the SQL query in the preceding item into relational algebra. (1 mark)

f) Recall the rewrite rules used in heuristics-based logical optimization discussed in the course unit. Cite two rewrite rules that you can apply to the relational-algebraic expression that constitutes your answer to the preceding item and explain the heuristic basis for applying each of them. (3 marks)

g) Again in the context of the suppliers, parts and catalogue relation schemas in the preceding items, assume that the data dictionary holds the following metadata:

<table>
<thead>
<tr>
<th>relation</th>
<th>cardinality</th>
<th>value distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>parts</td>
<td>100</td>
<td>V(parts, pid) = 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V(parts, colour) = 8</td>
</tr>
<tr>
<td>catalogue</td>
<td>200</td>
<td>V(catalogue, pid) = 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V(catalogue, cost) = 5</td>
</tr>
</tbody>
</table>

Recall that, for a relation R and an attribute A, |R| denotes the estimated cardinality of R and V(R, A) denotes the estimated number of distinct values that appear in the column A of R. Draw the operator tree for the following logical plan:

π pid (σ colour = 'c' (parts) ⊤ pid σ cost = 40 (catalogue))

and, given the statistics in the table above, and assuming that bag semantics is used and that the join is the natural join, annotate each node with an estimate of the cardinality of its output. (10 marks)

h) Briefly explain the three different kinds of parallel architectures that arise as result of whether primary and/or secondary memory is shared between processors. In your explanation, make sure that you clearly state where the interconnect lies in the architecture. (3 marks)
2. a) Consider again the suppliers, parts and catalogue relation schemas used in the preceding question. Let the parts relation be partitioned as follows: partition P1 contains parts whose colour is either 'a' or 'b', partition P2 contains parts whose colour is either 'c' or 'd', and partition P3 contains parts whose colour is either 'e' or 'f'.

i) Write the three algebraic expressions that correctly define each of the partitions (3 marks)

ii) Consider now the following SQL query over the distributed database that results from the partitioning described above:

```sql
SELECT colour, COUNT(*)
FROM parts
WHERE colour <> 'a'
GROUP BY colour
```

Briefly explain how this query could be evaluated in a map-reduce engine. (6 marks)

b) Consider the notion of data appliance described in this course unit.

i) Briefly explain the difference between a Type 0 and Type 1 data appliance. (1 mark)

ii) Briefly explain what is the general strategy in data appliances regarding the use of indexes. (2 marks)

iii) Briefly explain, in terms of join strategies, why data appliances tend to come with unusually large amounts of primary memory. (2 marks)
c) Assume that a sliding window equijoin with width = 6 is running, under bag semantics, over two streams $S_1(ts, a, b)$ and $S_2(ts, c, d, b)$, where $ts$, in both streams, is the implicit timestamp attribute. Let the current state of the windows be as follows:

$S_1 = [(1, x, 3), (2, y, 1), (3, t, 2), (4, u, 1), (5, s, 1), (6, b, 1)]$

$S_2 = [(1, r, u, 5), (2, b, v, 7), (3, k, u, 9), (4, d, u, 8), (5, e, u, 8), (6, h, v, 2)]$

Assume the query being executed is:

$$\text{SELECT * FROM } S_1 \text{ [RANGE NOW TO NOW-5], } S_2 \text{ [RANGE NOW TO NOW-5] WHERE } S_1.b > S_2.b;$$

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

i) At $ts = 7$, when the state is as above, $(7, e, 3)$ arrives in $S_1$

ii) At $ts = 8$, when the state is as resulted from the previous item, $(8, j, v, 1)$ arrives in $S_2$

(6 marks)

d) Consider the annotated routing tree below. Arrows denote child-to-parent communication links, nodes are labelled with node name, edges are labelled with a list denoting the values transmitted through that edge.

Draw an equally-shaped routing tree but label the edges under the assumption that one is doing a tree-staged $\text{SUM}$, rather than sending all the data back to the base station as in the figure above, then calculate the reduction (in percentage terms) obtained in terms of number of bytes transmitted by using tree-staged aggregation compared to the number of bytes transmitted in the figure above. (10 marks)