Exercise 9 “Implementation of Databases”

Due until 9 Jan, 2008 (before exercise) WM 07/08

Note: tasks with a ♠ are advanced and optional, though you are encouraged to do them.

Exercise 9.1[Selectivity Estimation] :
Suppose we have two relations $R(x, y)$ and $S(y, z)$. It is also known that $T(R) = 52$, $T(S) = 78$ and $V(R, y) = V(S, y) = 20$, where $T(R)$ is the number of tuples of relation $R$ and $V(R, A)$ is the number of distinct numbers of attribute $A$ in relation $R$. Furthermore, $\min(R, y) = \min(S, y) = 0$ and $\max(R, y) = \max(S, y) = 100$. Consider the following expressions:

- $\sigma_y=2(R)$
- $\sigma_y<3(R)$
- $R \bowtie S$

1. If we assume uniform distribution, what is the selectivity of each of the expressions? (3 pt.)

2. If the database holds a histogram of the two relations as follows, what are the selectivities?\(^1\)

<table>
<thead>
<tr>
<th>buckets(y)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>count(R)</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>count(S)</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>48</td>
</tr>
</tbody>
</table>

(3 pt.)

3. What are the error rates\(^2\) of the first estimation against the second? (3 pt.)

Exercise 9.2[Optimization Heuristics] :
The four heuristics below all have been used as ways to reduce the search space of query plans during query optimization. For each of the four heuristics:

- Push selection predicates down the plan.
- If a selection predicate is expensive to evaluate, pull it up the plan.
- Perform joins before cross-products.
- If both operands of a join are already sorted, use a sort-merge join.

1. Briefly state why applying the heuristic is generally expected to yield good plans. (An explanation consisting of a few well-chosen words is sufficient.) (4 pt.)

\(^1\)Hint: 1. assume tuples within each bucket are uniform; 2. when joining two buckets, each tuple from the smaller one always matches some tuples from the larger one.

\(^2\) $\frac{|\text{est}_1 - \text{est}_2|}{\text{est}_2} \times 100\%$
2. Argue that the heuristic can sometimes fail: Describe or show a query plan (logical or physical) and database characteristics (e.g., statistics, auxiliary structures, or anything else typically used in plan costing) such that applying the heuristic is very likely to result in a worse plan than not applying it. Give the simplest plan and characteristics you can come up with that contradicts the heuristic. (4 pt.)

Exercise 9.3[Query Optimization]:
Given are the relations r=(A,B,C), s=(C,D,E) and t=(E,F). Furthermore, the following statistics are given:

- \(|r| = 1000, |s| = 1500, |t| = 2000\)
- \(|\Pi_C(r)| = 700, |\Pi_C(s)| = 1100, |\Pi_E(s)| = 50, |\Pi_E(t)| = 250\)
- uniform distribution of all attribute values.

Consider the query \(r \bowtie s \bowtie t\).

1. Describe the two possible query plans to evaluate this query. (2 pt.)
2. Which query plan should be preferred to evaluate the query? Give reasons for your answer. (3 pt.)
3. If the memory is large enough to hold all relations and any intermediate results, which kind of join implementation would you use? (2 pt.)