CS4400: Database Systems
Homework 4
Cost of DB operations
Due February 27, 2017, 11:59pm

Instructions: Upload your submission as a PDF file on ICON under Assignments > Homework 4.

Goals for this assignment
- Get practice estimating the output size of queries and query plans
- Get practice estimating the cost of relational database plans

Complete the ICON Quiz "Cost of DB Operations" before attempting this assignment.

1. (2 points) "Database systems" Garcia-Molina, Ullman, Widom
Ex 16.4.1 (d) and (i), keeping in mind the that when we are only given T and V, we’ll assume the distribution of values in a particular column is uniform. Write down any other assumptions you make.

2. (4 points) Consider two relations R and S with the following statistics collected by the DBMS.

R(a, b)
V(R, a) = 50

<table>
<thead>
<tr>
<th>Value</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>[1-2]</td>
<td>100</td>
</tr>
<tr>
<td>[3-7]</td>
<td>100</td>
</tr>
<tr>
<td>Other</td>
<td>100</td>
</tr>
</tbody>
</table>

S(a, c)
V(S, a) = 20

<table>
<thead>
<tr>
<th>Value</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>other</td>
<td>200</td>
</tr>
</tbody>
</table>

Provide the best estimate you can for $T(R \bowtie S)$. (useful materials: lecture 2/17 or ch16.4)
3. (6 points) This problem is modified from “Database Systems” Garcia-Molina, Ullman, Widom, Ex 16.5.6

Suppose we wish to compute the expression
\[ \tau_b (R(a, b) \bowtie S(b, c)) \], that is, we join R and S on column b and then sort by column b.

Assume the following properties of plans explored by the optimizer
i. Join can be implemented with either a) two-pass merge sort on each relation followed by merge join (15.4.1 or lecture) or b) two-pass hash-join (15.5.5-7 or lecture)
ii. Any relation or result of an expression can be sorted by merge sort
iii. The result of a join can be passed to the next operator without being written temporarily to disk
iv. The result of a sort can be passed to the next operator without being written temporarily to disk
v. \[ B(R) = 1000, B(S) = 500, B(R \bowtie S) = 5000 \]

Assume that \( 1000 < M < 5000 \) (M is the number of blocks that can fit in memory at one time).

What subexpressions would a Selinger-style optimizer consider, and what is the cost (in terms of number of disk I/Os required) for each of those subexpressions? Show the costs both symbolically and numerically.

Reminders:
- The number of memory blocks is important because it tells you which version of an algorithm you can use.
- Be careful not to over count the number of disk I/Os across an entire plan. You should only count an I/O when a block is read from disk or written to disk.
- Sort-merge join treats both relations the same, but hash join uses one relation for building the hash table and one relation for probing.

Extra credit

(Up to 3 points) Only attempt the extra credit if you've finished the homework. Perform an interesting experiment with PostgreSQL's statistics collection. You should state a hypothesis, take measurements to test the hypothesis, and discuss the results. Documentation on PostgreSQL statistics is here: [https://www.postgresql.org/docs/9.1/static/monitoring-stats.html](https://www.postgresql.org/docs/9.1/static/monitoring-stats.html).