1 Short answer

Assuming $B(R) = 1000$, $T(R) = 200,000$ and $V(R, A) = 500$, estimate the number of disk I/O’s for an index-based selection for $\sigma_{A=55}(R)$ in each of the following cases:

i. (2 points) The system uses an unclustered index on $R.A$.

   Number of I/O’s:

   i. $\underline{400}$

ii. (2 points) The system uses a clustered index on $R.A$.

   Number of I/O’s:

   ii. $\underline{2}$

(2 points) The main advantage of a clustered index over an unclustered index is that the clustered index uses less space.

(i) $\underline{\text{False}}$

True or false?

(2 points) The main reason why we cannot create too many indexes is because they will slow down updates to the database.

(j) $\underline{\text{True}}$

True or false?
2 Cost Evaluation

Below, we consider possible execution plans for the following query, which returns the names of grocers that sell at least one very inexpensive ingredient:

```
SELECT gname
FROM Grocer G, Sells S
WHERE G.gid = S.gid AND S.price < 0.26
```

Assume the following statistics about the two relations:

<table>
<thead>
<tr>
<th>Table</th>
<th># tuples</th>
<th># blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grocer</td>
<td>1000</td>
<td>25</td>
</tr>
<tr>
<td>Sells</td>
<td>10,000</td>
<td>100</td>
</tr>
</tbody>
</table>

Assume the following statistics about the price column of S:

<table>
<thead>
<tr>
<th># distinct</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.01</td>
<td>2.50</td>
</tr>
</tbody>
</table>

At the bottom of this page, we provide some formulas from lecture that may be useful if you do not already have them in your own notes. The problems start on the next page.

Estimated cost of X JOIN Y:

- Using a nested loop, the cost is $B(X) + B(X) \times B(Y)$, where $B(X)$ is # blocks in $X$.
- Using a clustered index on $Y(A)$, the cost is $B(X) + T(X) \times B(Y) \times E$, where $T(X)$ is # tuples in $X$, and $E$ is the selectivity of the condition $A = c$.
- Using an unclustered index on $Y(A)$, the cost is $B(X) + T(X) \times T(Y) \times E$.

Estimated selectivity of conditions:

- For $A = c$, the selectivity is $1 / (#$ distinct values of $A)$
- For $A < c$, the selectivity is $(c - lowest \ value \ of \ A) / (highest - lowest \ value \ of \ A)$
1. The total cost of the plan below is ______370______ disk block I/Os.

![Diagram of a nested loop join plan]

Notes:
- The cost to scan \( S \) is \( B(S) = 100 \). The resulting table has 10 blocks since the selectivity of \( \text{price} < 0.26 \) is 1/10. It costs 10 to write those to disk.
- The cost of the join is then \( 10 + 10 \times B(G) = 260 \).
2. The total cost of the plan below is \(10,100\) disk block I/Os.

\[
\text{(on the fly)} \quad \pi \quad \text{gname} \\
\text{(on the fly)} \quad \sigma \quad \text{price} < 0.26 \\
\text{(indexed join)} \quad \nabla \quad \text{G.gid=S.gid} \\
\text{(file scan)} \quad S \quad G \quad \text{(file scan)}
\]

Notes:
- The cost of the indexed join is \(B(S) + T(S) \cdot B(G) \cdot E\), but \(E = 1/B(G)\) because a primary key lookup will always read just a single block (to find one tuple).
- Thus, the total cost is \(B(S) + T(S) = 100 + 10,000 = 10,100\).

3. Circle the plan that is probably faster in this case:

\[\text{nested loop join (1)} \quad \text{indexed join (2)}\]