CSE 132C Database System Implementation

Exercise 4

Time tip: Roughly 45sec to 1min per 1pt
## Common Info: Netflix Schema

### Ratings / R

<table>
<thead>
<tr>
<th>RatingID</th>
<th>Stars</th>
<th>RateDate</th>
<th>UID</th>
<th>MID</th>
</tr>
</thead>
<tbody>
<tr>
<td>7254</td>
<td>4.5</td>
<td>12/15/19</td>
<td>839</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Users / U

<table>
<thead>
<tr>
<th>UID</th>
<th>UName</th>
<th>Age</th>
<th>JoinDate</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Alvarez</td>
<td>39</td>
<td>11/02/14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Movies / M

<table>
<thead>
<tr>
<th>MID</th>
<th>Name</th>
<th>Year</th>
<th>Director</th>
</tr>
</thead>
<tbody>
<tr>
<td>492</td>
<td>Parasite</td>
<td>2019</td>
<td>Bong Joon-Ho</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|     |     |     |                      |
|     |     |     |                      |
Q1) [4 x 5pts] You are given two instances of R (R1 and R2) and the following statistics of the number of pages in each relation and the allotted buffer memory size in pages B. Page size is 8 KB. Suppose all attributes are 8 bytes long. Assume UID and MID are uniformly distributed in R. Ignore output write costs. What is the I/O cost (in number of pages) of the following operations using any of the implementations discussed in the lecture?

\[(N_{R1}, N_{R2}, B) = (40000, 25000, 5000)\]

A. Intersection of R1 and R2
B. Union of R1 and R2
C. Set difference R1 - R2
D. Set difference R2 - R1
Exercise

Q2) [3 x 4pts] You are given the following statistics of the number of pages in U and the allotted buffer memory size in pages B. Suppose all attributes are 8 bytes long, except U.Name, which is 40 bytes. Page size is 8 KB. You are also given a clustered AltRID B+ tree index on U with IndexKey (JoinDate, Age). RID length is also 8 bytes. What is the rough I/O cost (in number of pages) of the following operation with the specified implementation?

\( (N_U, B) = (10000, 500) \quad \gamma \text{COUNT(DISTINCT Age)}(U) \)

A. Hashing-based aggregate
B. Sorting-based aggregate
C. Index-based aggregate
Exercise

Q3) [3 x 4pts] You are given the following statistics of the number of pages in U and the allotted buffer memory size in pages B. Suppose all attributes are 8 bytes long, except U.Name, which is 40 bytes. Page size is 8 KB. You are also given a clustered AltRID B+ tree index on U with IndexKey (JoinDate, Age). RID length is also 8 bytes. What is the rough I/O cost (in number of pages) of the following operation with the specified implementation?

\[(N_u, B) = (10000, 500) \quad \gamma_{\text{JoinDate, AVG(Age)}}(U)\]

A. Hashing-based aggregate
B. Sorting-based aggregate
C. Index-based aggregate
Exercise

Q4) [3 x 4pts] You are given the following statistics of the number of pages in U and the allotted buffer memory size in pages B. Suppose all attributes are 8 bytes long, except U.Name, which is 40 bytes. Page size is 8 KB. You are also given a clustered AltRID B+ tree index on U with IndexKey (JoinDate, Age). RID length is also 8 bytes. What is the rough I/O cost (in number of pages) of the following operation with the specified implementation?

\[(N_u, B) = (10000, 7000) \quad \gamma \text{COUNT} (\text{DISTINCT Age})(U)\]

A. Hashing-based aggregate
B. Sorting-based aggregate
C. Index-based aggregate
Exercise

Q5) [3pts] Which of the following relational equivalencies hold?

A  $R \cup R = R$

B  $R \cap R = R$

C  $R \Join R = R$

D  All of A, B, C

E  None of the other options
Q6) [4pts] Which of the following relational equivalencies hold?

A  \[ \pi_A(R \times S) = \pi_{A \cap R.\star}(R) \times \pi_{A \cap S.\star}(S) \]

B  \[ \pi_A(R \Join S) = \pi_{A \cap R.\star}(R) \Join \pi_{A \cap S.\star}(S) \]

C  Both A and B

D  None of the other options
Exercise

Q7) [5pts] Which of the following relational equivalencies hold?

A \( \gamma_{A,COUNT(*)}(R) = \gamma_{A,COUNT(*)}(\pi_A(R)) \)

B \( \gamma_{MAX(B)}(R) = \gamma_{MAX(B)}(\pi_B(R)) \)

C \( \gamma_{SUM(B)}(R) = \gamma_{SUM(B)}(\pi_B(R)) \)

D All of A, B, C

E None of the other options
Exercise

Q8) [5pts] Which of these queries has/have at least 6 possible PQPs based only on the physical operators we saw in class?

A  \( \sigma_{\text{Stars} > 4}(R \bowtie M) \)

B  \( \pi_{\text{Name}}(\sigma_{\text{Stars} > 4}(R \bowtie M)) \)

C  \( \pi_{\text{Name}}(R \bowtie M) \)

D  All of A, B, C

E  None of the other options
Q9) You are given the following statistics of the number of pages of each relation in the Netflix database shown. Suppose all attributes are 8 bytes long, except U.Name, M.Name, and M.Director, each of is 40 bytes. Assume UID and MID are uniformly distributed in R. Ignore output write costs. Page size is 8 KB. What is the lowest estimate possible of the largest size of the output table of the following query (in # pages) with only the given information?

\[(N_R, N_U, N_M, B) = (80000, 20000, 5000)\]

A. [4pts] \( \pi UID, MID (R) \)

B. [5pts] \( \gamma Director, AVG(Stars) (M \bowtie R) \)

C. [6pts] \( R \bowtie U \bowtie M \)
Exercise

Q10) You are given the following statistics of the number of pages of each relation in the Netflix database shown and the allotted buffer memory size in pages \( B \). Suppose all attributes are \( 8 \) bytes long, except \( U\).Name, \( M\).Name, and \( M\).Director, each of is \( 40 \) bytes. Assume UID and MID are \textit{uniformly distributed} in \( R \). Ignore output write costs. Page size is 8 KB. No indexes exist in the database.

\[(N_R, N_U, N_M, B) = (50000, 10000, 2000, 5000)\]

A. [4pts] Which key-foreign key join in this database can NOT be executed using just one read of each base table?

B. [6pts] Propose a fully pipelined PQP for this query that can be executed with just one scan of each base table.

\[
\gamma \text{COUNT}(\ast) \left( \sigma \text{UID} = 123 \left( R \bowtie U \right) \right)
\]
Q10) (Continued)

$$(N_R, N_U, N_M, B) = (50000, 10000, 2000, 1000)$$

C. **[8pts]** What is the lowest possible I/O cost of this query using only the operator implementations discussed in the lectures? Consider possible algebraic rewrites too. Note the lower B. Explain your approach in detail.

$$\gamma_{Year, \text{COUNT}(\star)}(R \bowtie M)$$

D. **[10pts]** What is the lowest I/O cost of this query using only hash joins? Note the lower B. Explain your approach in detail.

$$R \bowtie U \bowtie M$$