CSE 132C
Database System Implementation

Arun Kumar

Review Discussion 1 with Answers
### 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>UNName</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>

---

Common Info: Netflix Schema
• All attributes in the given Netflix database are of fixed length and 8 bytes long, except for Director, Name, and UName, which are 40 bytes long.

• In the given instance, every UID in U and every MID in M arises at least once in R.

• The foreign keys have no dangling references.

• All tables are stored as heap files.

• The rough number of pages of R, U, and M are 100 million, 10 million, and 1 million, respectively.

• Page size is 4 KB (4000 bytes).
Q1) [3pts] Which 2-table key-foreign key join has the larger output in terms of cardinality? In terms of arity? In terms of number pages?

R JOIN U:
Cardinality: NTuples(R)
Arity: 5 + 4

R JOIN M:
Cardinality: NTuples(R)
Arity: 5 + 4

R JOIN M > R JOIN U in # pages
Q2) [5pts] What is the rough size in number of pages of the output of the 3-table natural star join in pages?

R JOIN U JOIN M

NPages(R) * (tuple size of output) / (tuple size of R)

Tuple size of R: 5 * 8 bytes -> 40 bytes

Output schema: all attributes of R; all attributes of U except UID; ... M ... -> 40 + 56 + 88 bytes

100 mil * (184 / 40) -> 460 mil
Q3) [5pts] What is the size (roughly) in number of pages of the output of the following query? Assume the aggregate attribute is 8 bytes long.

\[ \gamma_{UID, AVG(Stars)}(R) \]

Cardinality of result is # UIDs -> NTuples(U)

\[ \text{NPages(U)} \times \text{(tuple size of output)} / \text{(tuple size of U)} \]
\[ \rightarrow 10 \text{ mil} \times (16) / (64) \rightarrow 2.5 \text{ mil} \]
Q4) [4pts] Write a single relational algebra query to answer the following: Which user-movie ratings do not exist in R?

\[ \pi_{\text{UID}} (U) \times \pi_{\text{MID}} (M) \rightarrow \pi_{\{\text{UID, MID}\}} (R) \]
Exercise

Q5) [6pts] You are given 100,000 buffer frames. What is the lowest possible I/O cost in number of pages of sorting R on Stars using any of the optimizations discussed in the lectures?

Use internal replacement sort but no double buffering or blocked I/O

$$2N \times (1 + \lceil \log_{B-1} (\lceil N / (2\times B) \rceil) \rceil)$$

$N = 100 \text{ mil}; \quad B = 100 \text{ K}; \quad N / (2\times B) = 500$

$$2 \times 100 \text{ mil} \times (1 + \lceil \log_{99,999} (500) \rceil) \rceil)$$

$\rightarrow 200 \text{ mil} \times (1 + 1) \rightarrow 400 \text{ mil}$