Estimating result size

- Keep statistics for relation R
  - $T(R)$: # tuples in R
  - $S(R)$: # of bytes in each R tuple
  - $B(R)$: # of blocks to hold all R tuples
  - $V(R, A)$: # distinct values in R for attribute A
Size estimate for \( W = R_1 \bowtie R_2 \)

- Let \( x = \) attributes of \( R_1 \)
- \( y = \) attributes of \( R_2 \)

**Case 1**

\[ X \cap Y = \emptyset \]

\( S(W) \) same as \( R_1 \times R_2 \)

\( T(W) ? \)

**Case 2**

\( W = R_1 \bowtie R_2 \)

\[ X \cap Y = A \]

<table>
<thead>
<tr>
<th>R1</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2</td>
<td>A</td>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

**Assumption:**

\( V(R_1,A) \leq V(R_2,A) \Rightarrow \) Every A value in \( R_1 \) is in \( R_2 \)

\( V(R_2,A) \leq V(R_1,A) \Rightarrow \) Every A value in \( R_2 \) is in \( R_1 \)

“containment of value sets”
Computing \( T(W) \) when \( V(R1,A) \leq V(R2,A) \)

<table>
<thead>
<tr>
<th>R1</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R1</th>
<th>A</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Take 1 tuple

1 tuple matches with \( \frac{T(R2)}{V(R2,A)} \) tuples...

so \( T(W) = \frac{T(R2) \times T(R1)}{V(R2, A)} \)

\[ V(R1,A) \leq V(R2,A) \quad T(W) = \frac{T(R2) \times T(R1)}{V(R2,A)} \]

\[ V(R2,A) \leq V(R1,A) \quad T(W) = \frac{T(R2) \times T(R1)}{V(R1,A)} \]

[A is common attribute]
In general, \( W = R_1 \bowtie R_2 \)

\[
T(W) = \frac{T(R_2) \cdot T(R_1)}{\max\{ V(R_1, A), V(R_2, A) \}}
\]

Case 2 with alternate assumption

Values uniformly distributed over domain

This tuple matches \( T(R_2)/\text{DOM}(R_2, A) \) so

\[
T(W) = \frac{T(R_2) \cdot T(R_1)}{\text{DOM}(R_2, A)} = \frac{T(R_2) \cdot T(R_1)}{\text{DOM}(R_1, A)}
\]

Assume the same
Join: Tuple width

- In all cases:

$S(W) = S(R1) + S(R2) - S(A)$

size of attribute A

Size Estimation for Other Operations

- Projection: estimated size of $\Pi_A(r) = V(A, r)$
- Aggregation: estimated size of $\sum_A(r) = V(G, r)$
- Set operations
  - For unions/intersections of selections on the same relation: rewrite and use size estimate for selections
    - E.g., $\sigma_{\theta_1}(r) \cup \sigma_{\theta_2}(r)$ can be rewritten as $\sigma_{\theta_1 \lor \theta_2}(r)$
  - For operations on different relations:
    - estimated size of $r \cup s = \text{size of } r + \text{size of } s.$
    - estimated size of $r \cap s = \text{minimum size of } r \text{ and size of } s.$
    - estimated size of $r - s = r.$
  - All the three estimates may be quite inaccurate, but provide upper bounds on the sizes.
Note: for complex expressions, need intermediate T,S,V results.

E.g. \( W = [\sigma_{A=a} (R1) ] \bowtie R2 \)

Treat as relation \( U \)

\[ T(U) = T(R1)/V(R1,A) \quad S(U) = S(R1) \]

Also need \( V(U, \ast) \) !!

To estimate Vs

E.g., \( U = \sigma_{A=a} (R1) \)

Say \( R1 \) has attrs A,B,C,D

\( V(U, A) = \)
\( V(U, B) = \)
\( V(U, C) = \)
\( V(U, D) = \)
Example

R1

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>dog</td>
<td>1</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>dog</td>
<td>1</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>bat</td>
<td>1</td>
<td>50</td>
<td>10</td>
</tr>
</tbody>
</table>

\[
V(R1,A) = 3 \\
V(R1,B) = 1 \\
V(R1,C) = 5 \\
V(R1,D) = 3
\]

\[
U = \sigma_{A=a}(R1)
\]

\[
V(U,A) = 1 \\
V(U,B) = V(R1,B)
\]

\[
V(U,C) = \frac{T(R1)}{V(R1,A)}
\]

\[
V(D,U) \ldots \text{somewhere in between}
\]

Possible Guess

\[
U = \sigma_{A=a}(R)
\]

\[
V(U,A) = 1 \\
V(U,B) = V(R,B)
\]
For Joins  \[ U = R_1(A, B) \bowtie R_2(A, C) \]

\[
V(U, A) = \min \{ V(R_1, A), V(R_2, A) \} \\
V(U, B) = V(R_1, B) \\
V(U, C) = V(R_2, C)
\]

Example:

\[ Z = R_1(A, B) \bowtie R_2(B, C) \bowtie R_3(C, D) \]

- **R1**  
  \[ T(R_1) = 1000 \quad V(R_1, A)=50 \quad V(R_1, B)=100 \]

- **R2**  
  \[ T(R_2) = 2000 \quad V(R_2, B)=200 \quad V(R_2, C)=300 \]

- **R3**  
  \[ T(R_3) = 3000 \quad V(R_3, C)=90 \quad V(R_3, D)=500 \]
Partial Result: $U = R \Join S$

<table>
<thead>
<tr>
<th>$T(U)$</th>
<th>$V(U,A)$</th>
<th>$V(U,B)$</th>
<th>$V(U,C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1000 \times 2000 \div 200$</td>
<td>50</td>
<td>100</td>
<td>300</td>
</tr>
</tbody>
</table>

$Z = U \Join R3$

<table>
<thead>
<tr>
<th>$T(Z)$</th>
<th>$V(Z,A)$</th>
<th>$V(Z,B)$</th>
<th>$V(Z,C)$</th>
<th>$V(Z,D)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1000 \times 2000 \times 3000 \div 200 \times 300$</td>
<td>50</td>
<td>100</td>
<td>90</td>
<td>500</td>
</tr>
</tbody>
</table>
Estimation of Number of Distinct Values

Selections: $\sigma_\theta(r)$

- If $\theta$ forces $A$ to take a specified value: $V(A, \sigma_\theta(r)) = 1$.
  - e.g., $A = 3$
- If $\theta$ forces $A$ to take on one of a specified set of values: $V(A, \sigma_\theta(r)) = \text{number of specified values}$.
  - (e.g., $(A = 1 \text{ or } A = 3 \text{ or } A = 4)$),
- If the selection condition $\theta$ is of the form $A \text{ op } r$
  estimated $V(A, \sigma_\theta(r)) = V(A,r) \cdot s$
  where $s$ is the selectivity of the selection.
- In all the other cases: use approximate estimate of $\min(V(A,r), n_{\sigma_\theta(r)})$
  • More accurate estimates can be made using probability theory, but this one works fine generally

Estimation of Distinct Values (Cont.)

Joins: $r \bowtie s$

- If all attributes in $A$ are from $r$
  estimated $V(A, r \bowtie s) = \min(V(A,r), n_{r \bowtie s})$
- If $A$ contains attributes $A1$ from $r$ and $A2$ from $s$, then estimated $V(A, r \bowtie s) = \min(V(A1,r) \cdot V(A2,A1,s), V(A1 - A2,r) \cdot V(A2,s), n_{r \bowtie s})$
  • More accurate estimates can be made using probability theory, but this one works fine generally