Problem set 6,
Advanced Course on Databases

Group 2

July 25, 2021

Methodology

On this exercise Filipe Felício was the chairman. Filipe Felício and Luis Araújo since they live together had the opportunity work on this assignment side by side.

Exercises

1)

Assume that we have 3 tables r = (<u>A</u>, B, C), s = (<u>C</u>, D, E) and t = (<u>E</u>, F) where the primary keys of the relations are underlined. The table r has 1000 tuples, s has 1500 tuples and t has 750 tuples.

Estimate the size of the natural join $r \bowtie s \bowtie t$ and explain how you reason to calculate the estimate. Describe an efficient execution strategy for computing the join.

First, we join r and s and we joined them in C. C is a key in s and a foreign key in r, so each tuple of r joins exactly with one tuple in s:

$u = r \bowtie s$
$T(u) = T(r) = \ 1000 \ tuples$

After we join u with the table t on E, E is a key in t and a foreign key in u, so each tuple of u joins exactly with one tuple in t:

$T(u \bowtie t) = T(t) = \ 750 \ tuples$

2)

Assume we have the same tables r and s as in the previous question. We also know that the attribute B in the table r has a minimum value of 100 and a maximum value of 300. Transform the following expression into a more

efficient form and estimate the size of its result

First, we transform the expression into a more efficient form:

$$\sigma_{B\leq 250}\ (r\ \bowtie\ s)$$
$$= (\sigma_{B\leq 250}\ r)\ \bowtie\ s$$

Then, we estimate the size of its result:

$$T(\sigma_{B\leq 250}\ (r\ \bowtie\ s)) = T(r\ \bowtie\ s)\ \frac{250-100}{300-100} \equiv$$
$$\equiv T(\sigma_{B\leq 250}\ (r\ \bowtie\ s)) = T(r\ \bowtie\ s)\ \frac{150}{200} \equiv$$
$$\equiv T(\sigma_{B\leq 250}\ (r\ \bowtie\ s)) = 0.75 \times 1000 \equiv$$
$$\equiv T(\sigma_{B\leq 250}\ (r\ \bowtie\ s)) = 750$$

3)

Assume that we have 3 tables r = (<u>A</u>, B, D), s = (<u>C</u>, D, F) and t = (<u>E</u>, F) where the primary keys of the relations are underlined and the tables have the same number of tuples as in the previous questions. We also know the following about the number of distinct values of the attributes in the tables: V(B, r) = 200, V(D, s) = 300 and V(F, t) = 50.
Transform the following expression into a more efficient form and estimate the size of its result

$$\sigma_{A=1023}\ (r\ \bowtie\ s\ \bowtie\ t)$$

For a more efficient expression, we changed the expression to $\sigma_{A=1023}\ (r\ \bowtie\ (\ s\ \bowtie\ t))$. With this expression, the first join to compute is $s\ \bowtie\ t$ which is smaller then $r\ \bowtie\ s$. When intersecting s with t ($s \cap t$) we can find 50 tuples, and when intersection r with the result of the join of s with t ($r \cap (s \cap t)$, we end up with 300 + 50 = 350.

4)

Estimate the number of tuples that the following queries will produce. Please also explain how the estimates are calculated.

a) $\sigma_{dept="Physics"}\ (student)$

Number of tuples (nr) = 10000
V(dept,student) = 10
Estimated number of tuples = $\frac{nr}{V(dept, student)}$

. $= \frac{10000}{10}$

. $= 1000$

Review

This exercise was correct.

b) $\sigma_{credits \geq 120}\ (student)$

Number of tuples (nr) = 10000
v = 120
min(credits,student) = 0
max(credits,student) = 200

$$\text{Estimated number of tuples} = nr\ *\ \frac{v\ -\ min(credits, student)}{(max(credits, student)\ -\ min(credits, student))}$$

$$. \quad = 10000\ *\ \frac{120 - 0}{200 - 0}$$

$$. \quad = 6000$$

Review

Despite the fact that the solution indicates otherwise, this exercise is correct according to the material given:
If min(A,r) and max(A,r) are available in the metadata catalog we estimate c,the number of tuples satisfying the condition as:

- c = 0 $if\ v < min(A, r)$ (no tuples have a value smaller than the minimum)
- c = $nr\ if\ v \geq max(A, r)$ (all tuples have a value smaller than the maximum)
- c = $nr\ *\ \frac{v\ -\ min(credits, student)}{(max(credits, student)\ -\ min(credits, student))}$ otherwise

Considering that the last predicate is the one that verifies, the correct formula was chosen.

c) $\sigma_{year='2009'\ \wedge\ grade \neq 'F'}\ (takes)$

- **year = '2009'**
  Once the attribute "year" is a primary key the estimated number of tuples is 1.

- **grade ≠ 'F'**
  Number of tuples (nr) = 50000
  V(grade,takes) = 9

  $$\text{Estimated number of tuples} = nr\ -\ \frac{nr}{V(year, takes)}$$

  $$. \quad = 50000\ -\ \frac{50000}{9}$$

  $$. \quad = 44444,445$$

- **year = '2009' $\wedge$ grade $\neq$ 'F'**
  Estimated number of tuples = nr * $\frac{s1 * s2}{nr^2}$
  . $= 50000 * \frac{1 * 44444,45}{50000^2}$
  . $= 0,889$

Review

Here we forgot to divide n takes by V(year,takes) in the selection on year