Database Systems WS 2019/20
Prof. Dr.-Ing. Sebastian Michel
M.Sc. Nico Schäfer / M.Sc. Angjela Davitkova
Exercise 5: Handout 26.11.2019, Due 02.12.2019 16:00 MEZ https://dbis.cs.uni-kl.de

Question 1: Join-Size Estimation (1 P.)

a) Estimate the size of the join $R(a,b) \bowtie S(b,c)$ using histograms for $R.b$ and $S.b$. Assume $V(R,b) = V(S,b) = 20$ and the histograms for both attributes give the frequencies of the four most common values, as below, and further assume that every value appearing in the relation with the smaller set of values (R in this case) will also appear in the set of values of the other relation.

	0	1	2	3	others
$R.b$	5	4	10	5	36

	0	1	2	4	others
$S.b$	10	8	5	7	50

How does this estimate compare with the simpler estimate, assuming that all 20 values are equally likely to occur, with $T(R) = 60$ and $T(S) = 80$?

b) Estimate the size of the natural join $R(a,b) \bowtie S(b,c)$ if we have the following histogram information. Give a lower and upper bound for the join size and explain under which circumstances they appear.

	$b < 0$	$b = 0$	$b > 0$
R	400	100	200
S	400	300	800

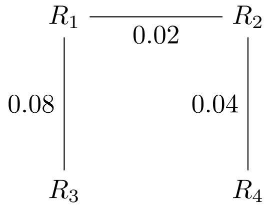
Question 2: Join-Ordering: Dynamic Programming (1 P.)

a) Manually create the DP-table for the relations A,B,C with cardinalities $|A| = 30$, $|B| = 50$, $|C| = 80$ and selectivities $f_{A,B} = 0.2$, $f_{B,C} = 0.4$ with C_{out} as cost function. Cross products are allowed this time. Please keep the replaced entries in the table and highlight the final ones.

b) Given the following DP-table with intermediate results and the query graph (with selectivities):

- $|R_1| = 20$
- $|R_2| = 10$
- $|R_3| = 30$
- $|R_4| = 10$

Relations	T	$C_{out}(T)$	$\|T\|$
$\{R_1, R_2\}$	$(R_1 \bowtie R_2)$	4	4
$\{R_1, R_3\}$	$(R_1 \bowtie R_3)$	48	48
$\{R_1, R_4\}$	$(R_1 \bowtie R_4)$	200	200
$\{R_2, R_3\}$	$(R_2 \bowtie R_3)$	300	300
$\{R_2, R_4\}$	$(R_2 \bowtie R_4)$	4	4
$\{R_3, R_4\}$	$(R_3 \bowtie R_4)$	300	300
$\{R_1, R_2, R_3\}$	$((R_1 \bowtie R_2) \bowtie R_3)$	13.6	9.6
$\{R_1, R_2, R_4\}$	$((R_1 \bowtie R_2) \bowtie R_4)$	5.6	1.6
$\{R_1, R_3, R_4\}$	$((R_1 \bowtie R_3) \bowtie R_4)$	528	480
$\{R_2, R_3, R_4\}$	$((R_2 \bowtie R_4) \bowtie R_3)$	124	120



Calculate the optimal bushy join tree for the relations $\{R_1, R_2, R_3, R_4\}$ with the DP-algorithm shown in the lecture.

Question 3: DP-Algorithm for chain queries (1 P.)

a) Write the pseudo code for a simple DP-algorithm which creates the optimal query tree for a chain query without cross products. The run-time has to be in $O(n^3)$. Bushy trees are allowed.

Submit the pseudo code and explain each line.

b) Implement this algorithm in a programming language of your choice and execute it on the following problem: 6 relations as chain R_1—0.1—R_2—0.7—R_3—0.2—R_4—0.3—R_5—0.4—R_6 with cardinalities of (from R_1 to R_6) 20, 10, 20, 20, 10, 20 and selectivities like shown in the chain.
Submit your code and your solution (HINT: You can modify the template of the previous sheet).

Question 4: Correctness of Unnesting (1 P.)

Provide unnested queries for the given nested queries. Show through an example, by specifying contents of tables and corresponding results, why the type of join (e.g,. INNER, LEFT OUTER) is important when unnesting a certain query. Considering the given queries, will the change in the type of the join impact the correctness of the query?

a)

```
SELECT DISTINCT P.playerId
FROM Player P
WHERE (
    SELECT COUNT(G.id)
    FROM Game G
    WHERE G.playerId = P.playerId
  ) >10
```

b)

```
SELECT DISTINCT P.name, (SELECT
   count(*)
FROM Game G
WHERE P.playerId = G.playerId)
FROM Player P
```