DB - HW12

16.5

Consider the relations r_1 (A,B,C) , r_2 (C,D,E) , and r_3 (E,F), with primary keys A, C, and E, respectively. Assume that r_1 has 1000 tuples, r_2 has 1500 tuples, and r_3 has 750 tuples. Estimate the size of $r_1 \bowtie r_2 \bowtie r_3$, and give an efficient strategy for computing the join.

Answer:

Size Estimation:

As the Natural Join operation is associative, we could estimate the size based on strategy $((r_1 \bowtie r_2) \bowtie r_3)$.

For C is the key of r_2, the result of $r_1 \bowtie r_2$ could at most include 1000 tuples.

Similarly, as E is the key of r_3, the result of $((r_1 \bowtie r_2) \bowtie r_3)$ could at most include 1000 tuples.

Thus, the size of $r_1 \bowtie r_2 \bowtie r_3$ would have no more than 1000 tuples.

Strategy:

1. Create index_1 on attribute 'C' for relation r_2 and index_2 on attribute 'E' for relation r_3
2. For each tuple in relation r_1:
   - Use index_1 to find the tuple in r_2 which matches the value of attribute 'C' o f relation r_1.
   - If found, index_2 to find the tuple in r_3 which matches the value of attribute 'E' o f relation r_1.

16.6

Consider the relation r_1 (A,B,C) , r_2 (C,D,E) , and r_3 (E,F) of Practice Exercise 16.5. Assume that there are no primary keys, except the entire schema. Let V(C, r_1), V(C, r_2) be 1100, V(E, r_2) be 50, and V(E, r_3) be 100. Assume that r_1 has 1000 tuples, r_2 has 1500 tuples, and r_3 has 750 tuples. Estimate the size of $r_1 \bowtie r_2 \bowtie r_3$ and give an efficient strategy for computing the join.

Answer:

Size Estimation:

For $r_1 \bowtie r_2$, it would have $\frac{n_{r1} * n_{r2}}{V(C, r_2)} = \frac{1000*1500}{1100} = \frac{15000}{11}$ tuples.

Rename the intermediate as r_1' , then $n_{r1'} = \frac{15000}{11}$.

Similarly, for $r_1' \bowtie r_3$, it would have $\frac{n_{r1'} * n_{r3}}{V(C, r_3)} = \frac{\frac{15000}{11} * 750}{100} = 10227$ tuples.

Strategy:

As $r_2 \bowtie r_3$ would have $\frac{n_{r2} * n_{r3}}{V(C, r_3)} = \frac{1500 * 750}{50} = 22500$ tuples which is much larger than the size of $r_1 \bowtie r_2$, thus we should do $r_1 \bowtie r_2$ first, and then make r_3 joined to the intermediate.

16.16

Suppose that a B^+-tree index on (*dept-name, building*) is available on relation *department*. What would be the best wat yo handle the following selection?

$$\sigma_{(bulding < 'Watson') \wedge (budget < 55000) \wedge (dept\_name = 'Music')}(department)$$

Answer:

As we have index on attributes *dept_name* and *building*, we can do the query like this:

1. Use the index on (*dept-name, building*) to locate the **first** tuple meeting the condition "building = 'Watson' AND dept_name = 'Music'".
2. Follow the pointers, retrieve all the successive tuples whose value of attribute 'building' is less than 'Watson'.
3. For every tuple retrieved in Step 2, check whether it the satisfies the condition "budget < 55000".

16.20

Explain how to use a histogram to estimate the size of a selection of the form $\sigma_{A \leqslant v}(r)$?

Answer:

Suppose the histogram H store the values in range $r_1,\ r_2,\ r_3,\ \ldots,\ r_n$,

and $r_i = [i\_low,\ i\_high],\ i \in\ [1, n],\ i\ \in\ \mathbb{Z}^+$.

- Case 1 —— $i\_high < v$

  Add the number of tuples = $H(r_i)$ to the estimated total.

- Case 2 —— $i\_low \leq v < i\_high$

  Assume that values in r_i are uniformly distributed:

  Add $\frac{v - r_{i\_low}}{r_{i\_high} - r_{i\_low}} * H(r_i)$ to the estimated total.