CS150A Homework 2 – Writing

School of Information Science and Technology
December 2, 2022

December 2, 2022

1 B+ tree Refinement (10 pts)

How many I/Os would it cost to insert an entry into our table if we had a height 3, unclustered alternative 3 B+ tree in the worst case? Assume that the cache is empty at the beginning and there are at most 6 entries with the same key. And what is the answer for a height 5, clustered alternative 3 B+ tree in the best case? Assume each page is at most 2/3 full.

12 7

Reference to Practice Questions in Note 4-B+Trees. In a clustered alternative 3 B+ tree, we need 6I/O to find the key. For the best case, the entry is in the first entry of the page, so we only need one more I/O to read the entry.

2 Relational Algebra (40 pts)

As shown in the following figures, there are three instances corresponding to three relations: Sailors, Reserves and Boats.

sid	*sname*	*rating*	*age*
22	Dustin	7	45.0
29	Brutus	1	33.0
31	Lubber	8	55.0
32	Andy	3	25.0
58	Rusty	10	35.0
74	Horatio	9	35.0

Fig. 1. An instance of Sailors

sid	*bid*	*day*
22	101	10/10/98
22	104	10/7/98
31	102	11/10/98
31	103	11/6/98
64	102	9/8/98
74	103	9/8/98

Fig. 2. An instance of Reserves

bid	*bname*	*color*
101	Interlake	blue
102	Interlake	red
103	Clipper	green
104	Marine	red

Fig. 3. An instance of Boats

Use relational algebra to describe the following queries.

1. Find the names of sailors who reserved boats after 10/7/98.
   $\pi_{sname}(Sailors \bowtie \sigma_{day \geq '10/7/98'}(Reserves))$
2. Find the age of sailors who have reserved boat Interlake with blue color.
   $\pi_{age}(Sailors \bowtie Reserves \bowtie \sigma_{bname='Interlake' \wedge color='blue'}(Boats))$
3. Find the name of boats which have been reserved by sailors rating 7 or higher.
   $\pi_{bname}(Boats \bowtie Reserves \bowtie \sigma_{rating \leq 7}(Sailors))$
4. Find the color of boats which once reserved by sailors Dustin on 10/10/98.
   $\pi_{day}(Boats \bowtie \sigma_{day='10/10/98'}(Reserves) \bowtie \sigma_{sname='Dustin'}(Sailors))$
5. Find the name of sailors who didn't reserved any boat named Marine.
   $\pi_{sname}(Sailors) - \pi_{sname}(Sailors \bowtie Reserves \bowtie \sigma_{bname='Marine'}(Boats))$

3 Joins (10 pts)

Determine whether each of the following statements is True or False.

a. Block Nested Loops join will always perform at least as well as Page Nested Loops Join when it comes to minimizing I/Os. True

b. Grace hash join is usually the best algorithm for joins in which the join condition includes an inequality. False

4 Query Optimization (40 pts)

Consider two relations R(a, b) and S(a), with 1000 tuples and 500 tuples respectively. We have an index on R.a with 50 unique integer values uniformly distributed in the range [1, 50], and an index on S.a with 25 unique integer values uniformly distributed in the range [1, 25]. We do not have an index on

R.b.
Use selectivity estimation to estimate the number of tuples produced by the following queries.

1. SELECT * FROM R
   No predicates, so select all tuples
   1000

2. SELECT * FROM R WHERE a = 99
   99 is out of the range for index on R.a distributed in the range [1, 50], so it won't have any output. Sel = 0
   0

3. SELECT * FROM R WHERE b = 99
   Sel = 1/10 because we do not have any information on b. We use 1/10 as default.
   1000 * (1/10) = 100

4. SELECT * FROM R WHERE a $\leq$ 24
   Sel = (v – low) / (high – low + 1) + 1 / distinct values = (24 – 1) / (50 – 1 + 1) + 1/ 50 = 12/25
   1000 * (12/25) = 480

5. SELECT * FROM R WHERE b $\leq$ 10
   Sel = 1/10
   1000 * (1/10) = 100

6. SELECT * FROM R WHERE NOT a $\leq$ 24
   Sel = 1- Sel(a $\leq$ 24) = 1- 12/25 = 26/50
   1000 * (26/50) = 520

7. SELECT * FROM R WHERE a $\leq$ 24 AND b $\leq$ 10
   Sel = Sel(a $\leq$ 24) * Sel(b $\leq$ 10) = 12/25 * 1/10 = 6/125
   1000 * (6/125) = 48

8. SELECT * FROM R WHERE a $\leq$ 24 OR b $\leq$ 10
   Sel = Sel(a $\leq$ 24) + Sel(b $\leq$ 10) – Sel(a $\leq$ 24) * Sel(b $\leq$ 10) = 12/25 + 1/10 - 6/125 = 133/250
   1000 * (133/250) = 532

9. SELECT * FROM R WHERE a = b
   We don't have information on b, so we just use the number of distinct values in a: Sel = 1/50
   1000 * (1/50) = 20

10. SELECT * FROM R, S WHERE R.a = S.a
    Sel = 1/max(50, 25) = 1/50
    (1000 * 500) * (1/50) = 10,000