CS150A Quiz 4

External Hashing and Sorting

Your "dream machine" allocates 32 MB for the buffer (memory) for its external hashing and sorting algorithms.

All input is read from disk, and all output is written to disk. The I/O cost is the number of page reads/writes, where each page is 128 KB large.

Note that 1024 KB = 1 MB.

1. **Q1: How many passes are required to fully sort a 8192 MB file with external merge sort?**

   $N = \frac{8192 \times 1024}{128} = 65536$ pages

   $B = \frac{32 \times 1024}{128} = 256$ pages

   # of passes $= \log_{B-1} \frac{N}{B} + 1$

   $= 3.$

2. **Q2: What's the I/O cost of fully sorting a 8192 MB file with external merge sort?**

   Cost $= 2N *$ (# of passes)

   $= 65536 \times 2 \times 3 = 393216$ I/O

3. **Q3: Suppose we double the size of our buffer, to 64 MB. What is the largest file size (in MB) that we can externally sort in two passes?**

   largest size (MB) = $B(B-1)$

   double size: 512 pages

   $= 512 \times 511 = 261632$ pages

   $= 32704$ MB

4. **Q4: Generalizing Q3, if we double the size of our buffer, approximately how much larger of a file can we externally sort in k passes?**

   Mark only one oval.

   ○ 2 times larger

   ○ 2^2 times larger

   ○ 2k times larger

   [✓ 2^k times larger]

   ○ k^2 times larger

   $B \to 2B.$

   k pass $= \log_{B-1} \frac{N}{B}.$

   — B

   $B^k \cdot B = N.$

   $(2B)^k \cdot 2B = N.$

For Q5 and Q6, use 32 MB for the size of the buffer.

5. **Q5: You decide to separate your 8192 MB dataset with external hashing. How does the I/O cost of externally hashing the file compare with the I/O cost of externally merge sorting the file?**

Assume that the data is uniformly distributed on the hashed key and that your hashing function distributes the records into partitions evenly.
Mark only one oval.

○ External merge sort will use fewer I/O's

☑ They have the same I/O cost

○ External hashing will use fewer I/O's

4*N I/Os.

✓ 6. **Q6: Suppose you are hashing a file and one of the partitions is 36 MB after the first pass (all other partitions can fit in the 32 MB buffer). How much larger (in I/Os) is the cost of externally hashing this file, compared to a scenario (with the same file) in which no partitions are ever oversized?**

Assume that a new hash function is chosen for the second pass such that the records are distributed in a way that guarantees subsequent partitions to be under 32 MB.

we can do partition on that 36 MB page without operation on others.

$\# \text{of pages} = \frac{36 \times 1024}{128} = 288$ pages.

$288 \times 2 = 576$ I/Os.

Relational Algebra

Given the following schema, let's look at some relational algebra!

```
Boats {
  bid int,
  color varchar(20),
  primarykey(bid)
}

Sailors {
  sid int,
  sname varchar(50),
  primarykey(sid)
}

Reserves {
  sid int,
  bid int,
  r_date char(10),
  primarykey(sid, bid, r_date),
  foreignkey(sid) references Sailors,
  foreignkey(bid) references Boats
}
```

Recall that π is project, σ is select, $\bowtie$ is join, and ρ is rename.

$A)\ \pi_{sname}(\sigma_{color = \text{'pink'}}(Reserves \bowtie Boats) \bowtie Sailors)$

$B)\ \pi_{sname}(\pi_{sid}(\sigma_{color = \text{'pink'}}(Reserves \bowtie Boats)) \bowtie Sailors)$

$C)\ \pi_{sname}(\sigma_{color = \text{'pink'}}(Reserves \bowtie Sailors) \bowtie Boats)$ ← color here

no color.

$D)\ \pi_{sname}(\sigma_{color = \text{'pink'}}(Reserves \bowtie Boats \bowtie Sailors))$

$E)\ \sigma_{color = \text{'pink'}}(\pi_{sname}(Reserves \bowtie Boats) \bowtie Sailors)$

$F)\ \sigma_{color = \text{'pink'}}(\pi_{sname}(Reserves \bowtie Sailors) \bowtie Boats)$

$G)\ \sigma_{color = \text{'pink'}}(\pi_{sname}(Reserves \bowtie Boats \bowtie Sailors))$

7. **Q7: Which of the relational algebra(s) above describe(s) the name of all sailors who have reserved pink boats?**

Check all that apply

Check all that apply.

☑ A
☑ B
☐ C
☑ D
☐ E
☐ F
☐ G

E, F, G } remove "colors" by early π (✓) Select.

8. **Q8: Which one of the above expressions that is correct, if executed as a query plan, is the most performant?**

↘ perform best

Mark only one oval.

○ A
[✓ B] → only join on the wanted col
○ C
○ D
○ E
○ F
○ G

π sid keeps no 3-table join

blue bid and, Boat Rserve ≥ 2016

$\rho(temp, \pi_{bid}(\sigma_{color = \text{'blue'}} Boats) \cap \pi_{bid}(\sigma_{r\_date \geq \text{'2016-01-01'}}(Reserves \bowtie Boats)))$

$\rho(result1, \pi_{sname}(Reserves \bowtie temp \bowtie Sailors))$

$\rho(temp1, \pi_{sid}(\sigma_{color = \text{'blue'}}(\sigma_{r\_date \geq \text{'2016-01-01'}}(\cancel{Reserves \bowtie Boats}))))$

$\rho(result2, \pi_{sname}(temp1 \bowtie Sailors))$

$result1 - result2$

same

key: Reserves ⋈ temp all these boat

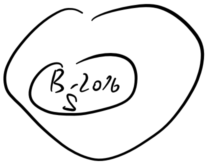
⋈ Sailors reserved. no 2016

all stricts: 2016, B.

9. **Q9: Which of the following *could* be in the output?** [illegible] hove!

Recall that B-A = {x: x in B and x not in A}. Choose all the options that could be in the set

Check all that apply.



☑ sailors who reserved blue boats but no other boats but before 2015

□ sailors who reserved boats in 2016 but only blue boats 2016 not Blue x . S. (B,2016)

□ sailors who have only reserved boats after 2016

☑ sailors who have reserved blue boats that were reserved by others in 2016 but not the sailor him/herself in 2016

$$\rho(temp1(sid1 \leftarrow sid), \pi_{color,sid}(Boats \bowtie Reserves))$$
$$\rho(temp2(sid2 \leftarrow sid), \pi_{color,sid}(Boats \bowtie Reserves))$$
$$\rho(temp3(sid3 \leftarrow sid), \pi_{color,sid}(Boats \bowtie Reserves))$$
$$\pi_{color}(\sigma_{(sid1 \neq sid2)}\sigma_{(sid1 \neq sid3)}\sigma_{(sid3 \neq sid2)}(temp1 \bowtie temp2 \bowtie temp3))$$

$S_1 \neq S_2 \neq S_3$

10. **Q10: What does the algebra above yield?**

Only one is correct

Mark only one oval.

● colors that have been chosen by at least three different sailors in their reservations

○ colors that have been chosen by at least three sailors in their reservations

○ the three most common boat colors among boats