CARNEGIE MELLON UNIVERSITY
COMPUTER SCIENCE DEPARTMENT
15-445/645 – DATABASE SYSTEMS (FALL 2019)
PROF. ANDY PAVLO

Homework #3 (by Erik Sargent)
Due: **Wednesday Oct 9, 2019 @ 11:59pm**

IMPORTANT:

- **Upload this PDF** with your answers to **Gradescope by 11:59pm on Wednesday Oct 9, 2019**.
- **Plagiarism**: Homework may be discussed with other students, but all homework is to be completed **individually**.
- **You have to use this PDF for all of your answers.**

For your information:

- Graded out of **100** points; **2** questions total
- Rough time estimate: ≈ 1 - 2 hours (0.5 - 1 hours for each question)

Revision : 2019/10/03 12:07

Question	Points	Score
Sorting Algorithms	40	
Join Algorithms	60	
Total:	100	

Number of Days this Assignment is Late:

Number of Late Day You Have Left:

Question 1: Sorting Algorithms . . . . [40 points]

We have a database file with six million pages (N = 6,000,000 pages), and we want to sort it using external merge sort. Assume that the DBMS is not using double buffering or blocked I/O, and that it uses quicksort for in-memory sorting. Let B denote the number of buffers.

(a) **[10 points]** Assume that the DBMS has five buffers. How many passes does the DBMS need to perform in order to sort the file?

□ 8 □ 10 ☑ 12 □ 14 □ 15

(b) **[5 points]** Again, assuming that the DBMS has five buffers. What is the total I/O cost to sort the file?

□ 72,000,000 □ 120,000,000 □ 132,000,000 ☑ 144,000,000 □ 168,000,000

(c) **[10 points]** What is the smallest number of buffers B that the DBMS can sort the target file using only two passes?

□ 50 □ 51 □ 52 □ 53 □ 172 □ 173 □ 174 ☑ 2,450 □ 2,451 □ 2,452 □ 3,000,000 □ 3,000,001

(d) **[10 points]** What is the smallest number of buffers B that the DBMS can sort the target file using only four passes?

□ 50 ☑ 51 □ 52 □ 53 □ 172 □ 173 □ 174 □ 2,450 □ 2,451 □ 2,452 □ 3,000,000 □ 3,000,001

(e) **[5 points]** Suppose the DBMS has ten buffers. What is the largest database file (expressed in terms of N, the number of pages) that can be sorted with external merge sort using five passes?

□ 89 □ 90 □ 91 ☑ 65,610 □ 65,611 □ 65,612 □ 590,488 □ 590,489 □ 590,490

Q1:

(a) $\left\lceil \frac{N}{B} \right\rceil = \frac{6{,}000{,}000}{5} = 1{,}200{,}000$ pages/run

$$1 + \left\lceil \log_{B-1} \left\lceil \frac{N}{B} \right\rceil \right\rceil = 1 + \lceil \log_{4} 1200000 \rceil$$

$$= 1 + 11 = 12$$

(b) $2 \cdot N \cdot (\#\text{ of passes}) = 2 \times 6{,}000{,}000 \times 12$

$$= 144{,}000{,}000$$

(c)

$$1 + \left\lceil \log_{B-1} \left\lceil \frac{N}{B} \right\rceil \right\rceil = 2$$

$$\Rightarrow \log_{B-1} \frac{N}{B} = 1 \Rightarrow B-1 = \frac{N}{B}$$

$$\Rightarrow 0 = B^2 - B - N$$

$$B = \frac{1 \pm \sqrt{1+4N}}{2} = \frac{1+\sqrt{1+4N}}{2}$$

$$= 2449.989 \approx 2450$$

$B = 2450 \Rightarrow \log_{2249} \lceil 2248.979 \rceil \approx 1$

(d) $1+\left\lceil \log_{B-1} \left\lceil \frac{N}{B} \right\rceil \right\rceil = 4$

$\log_{B-1} \frac{N}{B} = 3 \Rightarrow (B-1)^3 = \frac{N}{B}$

$\Rightarrow \quad B \cdot (B-1)^3 = N = 6000000$

$B=51, \quad \log_{50} 117648 = 2.98$

≈ 3

(e) $B=10$

$1+\left\lceil \log_9 \left\lceil \frac{N}{10} \right\rceil \right\rceil = 5$

$\Rightarrow \quad \log_9 \left\lceil \frac{N}{10} \right\rceil = 4 \Rightarrow \left\lceil \frac{N}{10} \right\rceil = 6561$

$\Rightarrow \quad N = 65610$

Question 2: Join Algorithms . . . . [60 points]

Consider relations `R(a, b)` and `S(a, c, d)` to be joined on the common attribute `a`. Assume that there are no indexes available on the tables to speed up the join algorithms.

- There are B = 36 pages in the buffer
- Table `R` spans M = 1800 pages with 100 tuples per page
- Table `S` spans N = 600 pages with 60 tuples per page

Answer the following questions on computing the I/O costs for the joins. You can assume the simplest cost model where pages are read and written one at a time. You can also assume that you will need one buffer block to hold the evolving output block and one input block to hold the current input block of the inner relation. You may ignore the cost of the writing of the final results.

(a) Hash join with `S` as the outer relation and `R` as the inner relation. You may ignore recursive partitioning and partially filled blocks.

i. **[5 points]** What is the cost of the partition phase?
□ 1,800 □ 2,400 □ 3,600 ☑ 4,800 □ 7,200

ii. **[5 points]** What is the cost of the probe phase?
□ 1,800 ☑ 2,400 □ 3,600 □ 4,800 □ 7,200

(b) **[10 points]** Block nested loop join with `R` as the outer relation and `S` as the inner relation:
□ 31,200 □ 31,800 □ 32,400 □ 33,000 ☑ 33,600

(c) **[5 points]** Block nested loop join with `S` as the outer relation and `R` as the inner relation:
□ 31,200 □ 31,800 □ 32,400 ☑ 33,000 □ 33,600

(d) Sort-merge join with `S` as the outer relation and `R` as the inner relation:

i. **[10 points]** What is the cost of sorting the tuples in `R` on attribute `a`? □ 3,600
□ 5,400 □ 7,200 □ 9,000 ☑ 10,800

ii. **[5 points]** What is the cost of sorting the tuples in `S` on attribute `a`? ☑ 2,400
□ 3,000 □ 3,600 □ 4,200 □ 4,800

iii. **[10 points]** What is the cost of the merge phase assuming there are no duplicates in the join attribute?
□ 1,200 □ 1,800 ☑ 2,400 □ 3,600 □ 4,800

iv. **[10 points]** What is the cost of the merge phase in the worst case scenario?
□ 2,400 □ 4,800 □ 600,000 ☑ 1,080,000 □ 1,200,000

↓
nested for loop

End of Homework #3

(a)

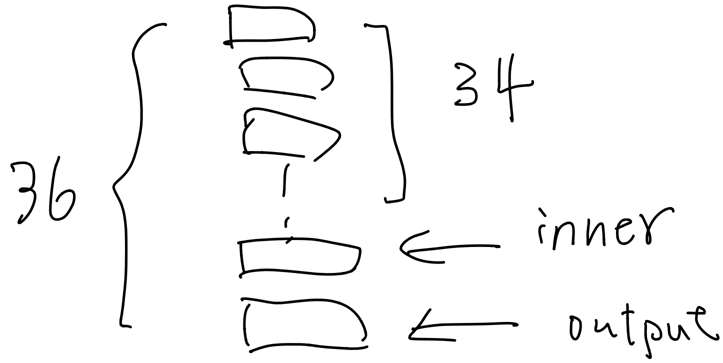
Partition:

$$2\cdot N + \overset{\text{read once write once}}{2}M = 2\cdot 2400 = 4800$$

Probe:

$$N+M = 2400$$

(b)



$$M + \left\lceil \frac{M}{B-2} \right\rceil \cdot N$$

$$= 1800 + \left\lceil \frac{1800}{34} \right\rceil \times 600$$

$$= 33600$$

(c)

$$N + \left\lceil \frac{N}{B-2} \right\rceil \times M$$

$$= 600 + \left\lceil \frac{600}{34} \right\rceil \times 1800 = 33000$$

(d)

(i) $2 \cdot M \cdot \left(1 + \left\lceil \log_{B-1} \left\lceil \frac{M}{B} \right\rceil \right\rceil \right)$

$= 2 \cdot 1800 \cdot \left(1 + \left\lceil \log_{35} \left\lceil \frac{1800}{36} \right\rceil \right\rceil \right)$

$= 3600 \cdot \left(1 + \left\lceil \log_{35} 50 \right\rceil \right)$

$= 3600 \cdot (1 + 2) = 10800$

(ii)

$2 \cdot N \cdot \left(1 + \left\lceil \log_{B-1} \left\lceil \frac{N}{B} \right\rceil \right\rceil \right)$

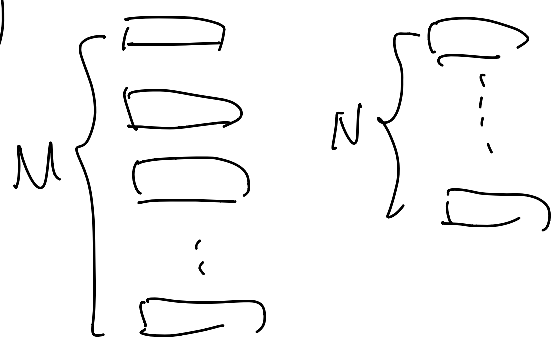
$= 2 \cdot 600 \left(1 + \left\lceil \log_{35} \left\lceil \frac{600}{36} \right\rceil \right\rceil \right)$

$= 1200 \cdot (1 + \log_{35} 17) = 1200 \times 2 = 2400$

(iii) No duplicates ⇒ $M + N$ using two pointers

$= 1800 + 600 = 2400$

(iv)



→ Block nested loop

$N + \left\lceil \frac{N}{B-2} \right\rceil \times M$

$= 600 + \left\lceil \frac{600}{34} \right\rceil \times 1800$

$= 33000$

⇒ Nested for loop w/o buffer

$M \cdot N = 1800 \times 600 = 1080000$