Name: Luis Castellanos

Student ID: 33489855

Purdue University (Fall 2025)
CS44000: Large-scale Data Analytics
Homework 1

IMPORTANT:

- Upload a pdf file with answers to Gradescope.
- Please use the either the latex template or word template to write down your answers and generate a pdf file.
  - Latex template: `https://www.cs.purdue.edu/homes/csjgwang/CS440/template.tex`
  - Word template: `https://www.cs.purdue.edu/homes/csjgwang/CS440/template.docx`

Problem	Score
1	
2	
3	
4	
5	
Total	

Problem 1 [20 points]

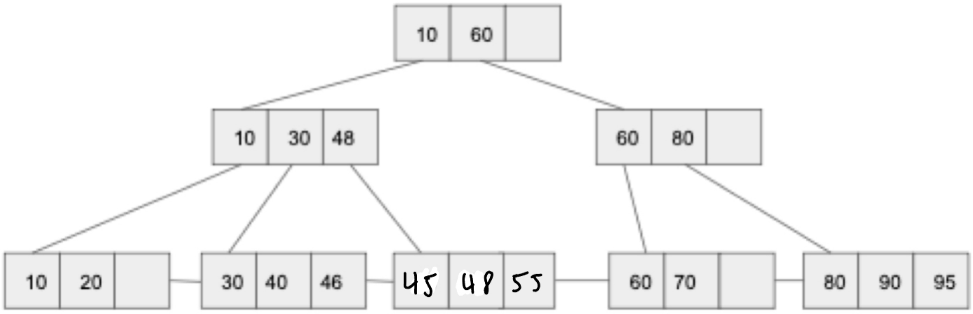
(a)

```
SELECT pid
FROM paper
WHERE pid NOT IN (
      SELECT pid FROM Reviews
);
```

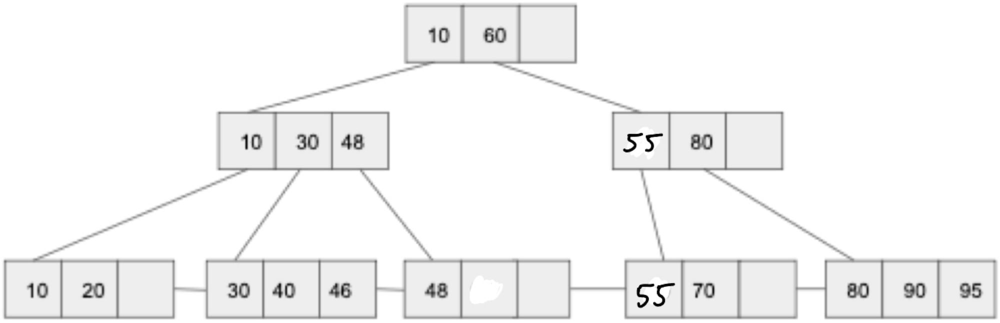
(b)

```
SELECT P.title
FROM Paper p
LEFT JOIN Reviews r ON p.pid = r.pid
GROUP BY p.pid, p.title
HAVING COUNT(r.rid) < 3:
```

Problem 2 [20 points]



(a)



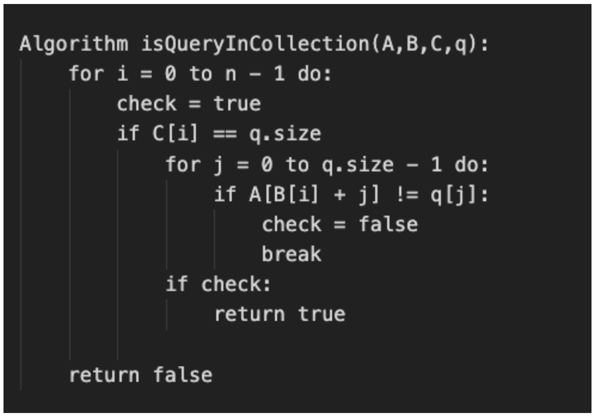
(b)

Problem 3 [20 points]

(a) External sorting benefits from having more memory by reducing the disk I/O operations. It then loads smaller chunks that fit in memory and sort the chunks individually, and merge them in multiple passes. When more memory is available larger chunks can be loaded and sorted so it decreases the number of merges and passes needed.

(b) Hash join performance improves with increased memory because it allows more data to be stored in the in memory hash table during the build phase, hence reducing the need for costly partition spills to disk and enabling faster probing.

Problem 4 [20 points]

(a) 

```
Algorithm isQueryInCollection(A,B,C,q):
    for i = 0 to n - 1 do:
        check = true
        if C[i] == q.size
            for j = 0 to q.size - 1 do:
                if A[B[i] + j] != q[j]:
                    check = false
                    break
            if check:
                return true

    return false
```

(b) Given that n is the number of elements in D and the cost of comparing two characters is $O(1)$, the time complexity of the algorithm is $O(n \cdot m)$. This is because the algorithm iterates over n strings, and for each, it checks every character with the query string q of length m when the lengths match, resulting in a total of $O(n \cdot m)$ comparisons at most.

Problem 5 [20 points]

(a) A declarative language allows users to specify what they want to retrieve from a database without detailing how to perform the operation, leaving the optimization and execution to the database management system. This approach offers improved productivity because users can focus on the desired outcome rather than complex logic, and enhanced performance, as the database engine can choose the most efficient execution.

(b) Rigorous two-phase locking ensures stronger consistency by holding all locks until the transaction commits, preventing cascading aborts and guaranteeing serializability, but it can lead to increased lock contention and reduced concurrency. Basic two-phase locking allows more flexibility by releasing locks earlier after the growing phase, thus improving concurrency, though it risks cascading aborts if a transaction rolls back after releasing locks.