Homework Assignment #2

(Due: Saturday, 25-Nov-2023, 11:59pm. Submit via Blackboard)

Question A. (35 marks)

1) What is a block anchor? How is it used in the implementation of a primary index? [4 marks]

2) Why can we have at most one primary or clustering index on a file, but several secondary indexes? [4 marks]

3) Consider the following *Movie* relation:

```
Movie (MovieID, Title, CategoryID, DirectorID, Rating)
```

MovieID is the primary key of the *Movie* relation. The *CategoryID* field contains 10 distinct values, and the records are evenly distributed among these values. The *DirectorID* field contains 200 distinct values, and the records are evenly distributed among these values. The *Movie* relation contains N = 4,000 records, with each record occupying 100 bytes. The file is sorted by the *CategoryID* field, and stored on a disk with the following configuration:

- Block size = 600 bytes
- Block pointer size = 6 bytes

Three indexes have been created on the *Movie* relation:

- A *clustering index* on the *CategoryID* field (Integer, 4 bytes)
- A *secondary index* on the *DirectorID* field (Integer, 4 bytes)
- A *bitmap index* on both *CategoryID* and *DirectoryID* fields (1 byte = 8 bits; each bitmap is stored as a *fixed-length record*)

Consider the following SQL query:

```
SELECT * FROM Movie WHERE CategoryID = 3 AND DirectorID = 16;
```

Assume that 4 records are generated as the result of this SQL query. Which index is the most efficient for answering this query? Justify your answer by comparing the estimated query cost (i.e., the number of block accesses). [27 marks]

Question B. (35 marks)

1) Consider two relations R(A, B) and S(A, C). Field A is the primary key of relation S, and the foreign key of relation R referencing S. Assume that *page-oriented nested-loop join* is used for R * S (natural join). Let R be the outer relation, and S be the inner relation. Given the following tuples of R and S, what are the first 4 tuples produced by R * S? Assume each block can store only 3 R tuples or S tuples. [8 marks]

- 1st tuple:
- 2nd tuple:
- 3rd tuple:
- 4th tuple:

A	B
7	x
2	z
9	y
8	y
3	w
9	x
1	w
3	y
5	z

R

A	C
8	1
4	3
2	6
1	5
3	7
5	1
7	8
9	2

S

2) Assume that R and S are stored on a disk with block size = 1000 bytes. Relation R contains 100,000 records, with each record occupying 20 bytes. Relation S contains 50,000 records, with each record occupying 50 bytes. Estimate the number of block accesses required for R * S using page-oriented nested-loop join. [6 marks]

3) Can we improve the efficiency of R * S by changing the join order, i.e. S as the outer relation and R as the inner relation? Justify your answer. [3 marks]

4) Consider the following relations in a *Company* database:

```
Employee (EID, EName, Email, Age, Address, Salary)
Department (DID, DName)
Join (EID, DID, Year)
```

EID is the primary key of *Employee*, and *DID* is the primary key of *Department. (EID, DID)* is the primary key of Join, where *EID* is the foreign key referencing *Employee* and *DID* is the foreign key referencing *Department. DName* is a unique field in *Department* relation. Given the following SQL query:

```
SELECT E.EName, E.Email, E.Salary
FROM Employee E, Join J, Department D
WHERE E.EID = J.EID AND D.DID = J.DID
        AND D.DName = 'Retailer' AND J.Year > 2016;
```

Show the initial query tree generated by the *conceptual evaluation strategy*. [5 marks]

5) Show the most efficient query tree after applying all the five steps of heuristic-based *query tree optimization*. [13 marks]

Question C. (30 marks)

1) *Two-phase locking* (2PL) is widely used for concurrency control in a relational database system. Consider the following schedule for two transactions T1 and T2:

```
S: W1(X); R2(Y); R1(Y); R2(X); C1; C2
```

"C1" means transaction T1 commits. For each of the following 2PL protocols: (1) Briefly describe the protocol; (2) State if the protocol allows schedule S, i.e., allows the actions to occur in exactly the order shown in schedule S; (3) Explain the reason why schedule S is allowed or not allowed under the protocol. [15 marks]

- Basic 2PL
- Conservative 2PL
- Strict 2PL

2) Consider two transactions T1, T2 and two data items X, Y.

```
T1: R(X); W(X); R(Y); W(Y)
T2: R(X); W(X)
```

Given the following schedule of interleaved operations from T1 and T2:

```
S: R1(X); W1(X); R2(X); R1(Y); W2(X); C2; W1(Y); A1
```

"A1" means transaction T1 aborts, and "C2" means transaction T2 commits. Answer each of the following questions. [15 marks]

- What type of *anomaly* occurs in this schedule S? Explain your answer.
- For each of T1 and T2, insert all the lock and unlock operations to make the transaction satisfy the strict 2PL protocol.
- Show that modifying the schedule according to strict 2PL can prevent the anomaly. Justify your answer.

Student Name: ______________________ Student ID: ______________________

Provide your answers for Question A on this page.

Provide your answers for Question B on this page.

Provide your answers for Question C on this page.