CS 564 Final Exam Spring 2018
Answers

A: Relational Algebra, SQL & Normalization **[22pts]**

I. **[5pts]** Consider a relational schema with two relations, $R(A, B)$ and $S(B, C, D)$, and the following query in Relational Algebra:

$$q = \pi_A(R \bowtie \sigma_{C=10}(S))$$

Which of the following queries are equivalent to q? Clearly **circle** all the correct options and only the correct options.

(a) $\pi_A(\sigma_{C=10}(R \bowtie S))$

(b) $\pi_A(R) \bowtie \pi_A(\sigma_{C=10}(S))$

(c) $\pi_A(\pi_B(R) \bowtie \pi_B(\sigma_{C=10}(S)))$

(d) $\pi_A(R \bowtie S) - \pi_A(R \bowtie \sigma_{C\neq 10}(S))$

(e) $\pi_A((R \bowtie S) - (R \bowtie \sigma_{C\neq 10}(S)))$

ANSWER: (a), (e)

II. **[9pts]** Consider a relation $R(A, B, C, D, E)$ with the following functional dependencies:

$$A \rightarrow B, C \qquad\qquad C \rightarrow D$$

For every one of the following decompositions, write YES/NO to answer whether it is lossless-join, dependency-preserving, or the result of a BCNF decomposition.

	lossless-join	**dependency-preserving**	**result of BCNF**
ABCD, DE			
ABC, ADE			
ABC, CD, AE			

ANSWER:

	lossless-join	**dependency-preserving**	**result of BCNF**
ABCD, DE	NO	YES	NO
ABC, ADE	YES	NO	YES
ABC, CD, AE	YES	YES	YES

III. **[8pts]** Consider the following schema:

Customer (cid, firstname, lastname, email)

Booking (bid, date, cid, tid, partySize)
Booking.cid is a foreign key referring to Customer.cid.

Suppose we want to ask the following SQL query:

```
SELECT C.lastname
FROM Customer C
WHERE EXISTS (
    SELECT *
    FROM Booking B
    WHERE B.partySize > 10 AND B.cid = C.cid);
```

Write an equivalent SQL query that does **not** use nesting or the WITH clause.

ANSWER:

B: Storage and Indexing **[28pts]**

I. **[5pts]** Suppose we have a table **Products** with the following attributes:

- ProductID INT PRIMARY KEY
- ProductName CHARACTER(255)
- ProductDescription CHARACTER(255)
- Code INT

Suppose that we are given a query workload where the **Code** attribute is used both in selection conditions and joins, while the **ProductID** column is not. What indexes should we have on this table? Clearly **circle** the correct option.

(a) CLUSTERED INDEX ON ProductID; CLUSTERED INDEX ON Code

(b) UNCLUSTERED INDEX ON ProductID; CLUSTERED INDEX ON Code

(c) CLUSTERED INDEX ON ProductID; UNCLUSTERED INDEX ON Code

(d) UNCLUSTERED INDEX ON ProductID; UNCLUSTERED INDEX ON Code

ANSWER: (b)

II. **[5pts]** Consider the relation $R(A, B, C, D)$ and suppose we want to evaluate the following selection predicate:

$$((A = 10) \text{ OR } (A > 10)) \text{ AND } (C = 10)$$

Which of the following indexes matches the predicate? Clearly **circle** all the correct options and only the correct options.

(a) hash index on (A)

(b) hash index on (C)

(c) hash index on (C, A)

(d) B+ tree index on (A, C)

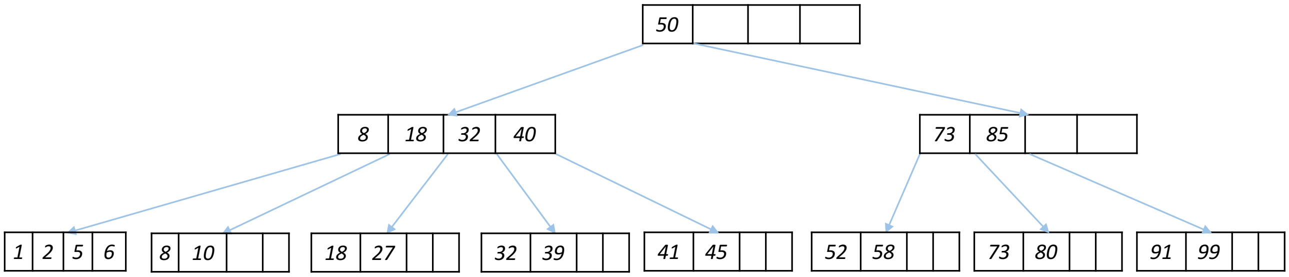
(e) B+ tree index on (C, B)

ANSWER: (b), (d), (e)

III. **[5pts]** Consider a relation $R(A, B, C, D)$ with $1,000,000$ tuples. Suppose that attribute A can take 10 distinct values, and each value needs 15 bits to be represented. Which one between a *bitmap* and *bitslice* index on A needs less space? Explain your answer in detail.

ANSWER: The bitmap index needs 11 bits per record, while the bitslice index needs 16 bits per records. So the bitmap index takes less space.

IV. **[13pts]** Consider the following B+ tree that has order $d = 2$ (so every node can hold at most 4 search key values):



(a) **[10pts]** Draw the B+ tree that results from inserting a data entry with key 3.

(b) **[3pts]** How many page reads and page writes does the insertion require?

ANSWER: 3 reads and 5 writes

C: Operator Implementation **[20pts]**

I. **[10pts]** We are given a relation R with $N = 500$ pages that we want to sort. The buffer pool has size $B = 20$. Fortunately, the first 300 pages of the relation are already sorted for us (that is, they form a sorted run). **What is the I/O cost of sorting the relation using the external sort algorithm?** Explain your answer in detail.

Assume that we do not use replacement sort for any initial sorting. You should include in your computation the cost of writing the final sorted result to disk.

ANSWER: In pass 0, we only need to create sorted runs for 200 pages, so we need $200 + 200 = 400$ **I/Os. Then, we need one more pass to merge (there are 10+1 runs, so they fit in the buffer). Total cost** $= 400 + 2 \times 500 = 1400$ **I/Os.**

II. **[10pts]** We are given two relations: R with 1,000 pages and S with 2,000 pages. We are performing a key-foreign key join of R and S wherein S has the foreign key attribute. Find values of the buffer pool size B where SMJ has a smaller I/O cost than BNLJ, and where BNLJ has a smaller I/O cost than SMJ. Explain your answer in detail.

	SMJ > BNLJ	**SMJ < BNLJ**
buffer size		

	SMJ > BNLJ	**SMJ < BNLJ**
buffer size	1100	240

D: Query Optimization **[20pts]**

Consider the following database schema:

Company (cid, cname, location)

Employee (eid, ename, age, cid)
Employee.cid is a foreign key referring to Company.cid

Product (pid, pname, price, cid)
Product.cid is a foreign key referring to Company.cid

Company has 1,000 tuples, and each record is 50 bytes long. Employee has 200,000 tuples, and each record is 20 bytes long. Product has 5,000 tuples, and each record is 20 bytes long. Each page can hold 5,000 bytes. The buffer pool has 102 frames.

I. **[10pts]** Consider the following SQL query:

```
SELECT  COUNT(DISTINCT E.eid)
FROM    Employee E, Company C
WHERE   E.cid = C.cid ;
```

Propose the best possible physical plan that evaluates the above query and compute its I/O cost. Explain your answer in detail.

ANSWER: The cost is only 800 I/Os, since the join can be avoided, and the count can de done without the distinct.

II. **[10pts]** Consider the following SQL query:

```
SELECT  SUM(price)
FROM    Employee E, Company C, Product P
WHERE   E.cid = C.cid AND P.pid = C.cid;
```

Propose the best possible physical plan for the above query that uses only BNLJ as a join algorithm and compute its I/O cost. (Hint: you may want to use pipelining). Explain your answer in detail.

ANSWER: We first join Product and Company (any can be the outer relation, since both fit together). Then we join with Company (intermediate result is the outer relation). The I/O cost is $800 + 10 + 20 = 830$**.**

E: TRANSACTION MANAGEMENT **[10pts]**

I. **[4pts]** For the following questions, **clearly circle** either True or False.

1. Suppose transaction T_1 has an S lock on tuple t. If transaction T_2 attempts to get an S lock on t, it will be successful.

   TRUE

2. The WAL protocol writes all modified pages to disk before a commit.

   FALSE

II. **[6pts]** Consider the following two transactions:

$$T1 : W(A), R(C), W(A)$$
$$T2 : R(B), R(A)$$

and the following interleaved schedule of the two transactions:

$$W_{T1}(A), R_{T2}(B), R_{T1}(C), W_{T1}(A), R_{T2}(A), Commit_{T1}, Commit_{T2}$$

Is this schedule serializable or not? If it is serializable, provide the equivalent serial schedule. Explain your answer in detail.

ANSWER: Yes, it is. First T_1, then T_2.