Monsoon Semester Examination, Session 2023-24

Examination & Semester: **M.Tech (Computer Science & Engineering) I Semester**
Subject: **Advanced DBMS (CSC502)**
Time: **3 Hours**
Max. Marks: **100**
Instructions:

(a) Answer the questions serially and to the point.
(b) Assume a suitable value against any missing data.

Section A (Marks: 50) – All Questions are COMPULSARY

Q. No.	Question	Marks
1. (a)	Consider the following relations: *Customer (CN, CS, CC); Account (AN, BN, BAL); Loan (LN, BN, AMT); Depositor (CN, AN); Borrower (CN, LN)* Write a query using relation algebra to "Find the names of all customers who have a loan at the bank." Rewrite the query to include not only the name, but also the city of residence for each customer. It has been observe that now one of the customer, say "Amit", no longer appears in the result, even though Amit does in fact have a loan from the bank. (i) Explain why Amit does not appear in the result. (ii) Suppose that you want Amit to appear in the result. How would you modify the database to achieve this effect? (iii) Again, suppose that you want Amit to appear in the result. Rewrite a query so that it accomplishes this desire without having to modify the database.	5
(b)	Use Armstrong's axioms to prove the soundness of the decomposition rule. Explain how functional dependencies can be used to indicate the following [relation schema is provided in 1(a)]: (i) A one-to-one relationship set exists between entity sets Account and Customer. (ii) A many-to-one relationship set exists between entity sets Account and Customer.	5
(c)	Suppose you are given a relation R with four attributes ABCD. For each of the following sets of FDs, assuming those are the only dependencies that hold for R, do the following: (a) Identify the candidate key(s) for R. (b) Identify the best normal form that R satisfies (1NF, 2NF, 3NF, or BCNF). (c) If R is not in BCNF, decompose it into a set of BCNF relations that preserve the dependencies. Show each decomposition step (if any) clearly. (i) $C \rightarrow D, C \rightarrow A, B \rightarrow C$ (ii) $ABC \rightarrow D, D \rightarrow A$	5 ABCD
(d)	Answer the following, with proper justification (Assume the height numbering starts from 0.): (i) What is the maximum fan-out of a B+ tree of an order x+1? (ii) What is the maximum number of keys an order 2 B+ tree with height 2 can store? (iii) For a height 3 B+ tree (leaf pages hold pointers to the corresponding records), how many I/Os are required for an equality search on the index key given that the index key is a primary key?	5
2. (a)	Due to hot weather, our machine has forgotten how to sort. Luckily, we have indexes on the join columns between R and S. R's index is an un-clustered index with a height of 2 and contains 10 leaf pages. S's index is a clustered index with a height of 2 and contains 12 leaf pages. R consists of 30 pages with 4 records on each page and S consists of 40 pages with 5 records on each page. What is the average cost of performing Sort Merge Join on R and S? You can assume that resetting to a record will never require accessing a previous leaf page or a previous data page and that there are enough memory pages to satisfy any operation.	5
(b)	Consider a relation 's' over the attributes A and B with the following characteristics: • 7,000 tuples with 70 tuples per page • A hash index on attribute A • The values that the attribute A takes in relation s are integers that are uniformly distributed in the range 1–200. (i) Assuming that the aforesaid index on A is unclustered and it requires 2 I/Os to locate right bucket, estimate the number of disk accesses needed to compute the query $\sigma_{A=18}(s)$. (ii) What would be the cost estimate if the index were clustered and it requires 4 I/Os to locate right bucket?	5
(c)	How does the quorum consensus protocol works? Can we say that the quorum consensus protocol is a generalization of biased and majority based protocol? Explain the same with a suitable example.	5
3. (a)	Briefly explain the concepts of Two-level serializability in the multi-database systems. Also emphasize on the restrictions that can be made for the better transaction management in the multi-database systems.	5
(b)	Explain, with a suitable example, the concepts of type and table inheritance of object-based databases.	5
(c)	NoSQL do not offer support for ACID properties of transactions. How the databases like MongoDB and Cassandra manage the concurrent transactions without the ACID properties? Justify your answer.	5

Section B (Marks: 30) – Attempt Any TWO

4. (a) Show the steps to remove extraneous attributes from the functional dependency. Find the canonical cover of the given F.

$F = \{A \rightarrow BC, B \rightarrow CE, A \rightarrow E, AC \rightarrow H, D \rightarrow B\}$ 5

(b) The keys 1, 3, 12, 4, 25, 6, 18, 20, and 8 are inserted in to an empty hash table of length 10 using open addressing with hash function as $H(i) = i^2 \bmod 10$ and linear probing. What is the resultant hash table and find the maximum probe value? 4

(c) Consider these relations with the following properties:
r(A, B, C) - 30,000 tuples and 25 tuples fit on 1 block
s(C, D, E) - 60,000 tuples and 30 tuples fit on 1 block

(i) Estimate the number of disk block accesses required for a natural join of r and s using a nested-loop join if r is used as the outer relation.

(ii) Estimate the number of disk block accesses required for a natural join of r and s using a block nested-loop join if s is used as the outer relation. Assume that there are more than 2000 memory buffers available to facilitate this operation, where each memory buffer can buffer one disk block. 6

5. (a) Briefly explain the different forms of parallelism. What form of parallelism is likely to be the most important for each of the following tasks [also justify your claim]:

(i) Increasing the throughput of a system with many small queries.

(ii) Increasing the throughput of a system with a few large queries, when the number of disks and processors is large. 5

(b) Design an efficient query plan for the following distributed query:
An application at site B wants to compute a join (Student $\bowtie_{Id=StudId}$ Gradesheet), where Student(Id, Major) is at site B and Gradesheet(StudId, CrsCode) is at site C. The result should be returned to B. Assume that semijoin is not used.
Also assume that

- The various lengths are:
  - Id and StudId are 8 bytes long;
  - Major is 3 bytes long;
  - CrsCode is 6 bytes long.
- Student has 15,000 tuples.
- 6,000 students are registered for at least one course. On the average, each student is registered for 5 courses. 4

(c) Consider the two-phase commit (2PC) protocol with write-ahead undo + redo logging [only the log file needs to be flushed to disk at the time of transaction commit]. Assume we have a system where the only failures involve hosts halting with their disks and logs intact, followed (potentially) by reboots, with no network message loss.

(i) Suppose there is a coordinator C and two participants P1 and P2. Suppose we have the following sequence of events in a 2PC:

C sends Prepare Transaction T1 to P1, P2
P1 sends Prepared to C
P2 sends No to C

Write an appropriate next message following 2PC, ignoring potential redundant retransmissions.

(ii) Suppose we have the following sequence of events in a 2PC:

C sends Prepare Transaction T2 to P1,P2
P1 sends Prepared to C
P2 sends Prepared to C
P2 crashes
P2 comes back online
C sends Commit Transaction T2 to P1, P2
P1 commits, sends DONE to C

Assuming standard 2PC operation in the face of the above scenario, what we can expect P2 to do next? 6

6. (a) Explain the following with respect to Cassandra database:

(i) Define and create a keyspace with replication.

(ii) Explain the concepts of partition and clustering key. Demonstrate the same with a suitable example. 5

(b) What is the need of a persistent programming language? Explain the different ways of making an object persistent. 4

(c) What is a need of TP monitor? Draw and explain the detailed structure of TP monitor.

Section C (Marks: 20) – Attempt Any TWO

7. (a) Consider the relations $r_1(A, B, C)$, $r_2(C, D, E)$, and $r_3(E, F)$, with primary keys A, C, and E respectively. Assume that r_1 has 1000 tuples, r_2 has 1500 tuples, and r_3 has 750 tuples. Estimate the size of $r1 \bowtie r2 \bowtie r3$, and give an efficient strategy for computing the join. 4

	(b)	Write down the sequence of steps (along with the name of the equivalence rules used) needed to transform $\pi_A((R \bowtie_{B=C} S) \bowtie_{D=E} T)$ into $\pi_A((\pi_E(T) \bowtie_{E=D} \pi_{ACD}(S)) \bowtie_{C=B} R)$. List the attributes that each of the schemas R, S, and T must have and the attributes that each (or some) of these schemas must not have in order for the above transformation to be correct.	6
8.	(a)	Consider a parallel DBMS in which each relation is stored by horizontally partitioning its tuples across all disks. *Employees(<u>eid: integer</u>, did: integer, sal: real)* *Departments(<u>did: integer</u>, mgrid: integer, budget: integer)* The *mgrid* field of Departments is the *eid* of the manager. Each relation contains 20-byte tuples, and the *sal* and *budget* fields both contain uniformly distributed values in the range 0 to 1,000,000. The Employees relation contains 100,000 pages, the Departments relation contains 5,000 pages, and each processor has 100 buffer pages of 4,000 bytes each. The cost of one page I/O is t_d, and the cost of shipping one page is t_s; tuples are shipped in units of one page by waiting for a page to be filled before sending a message from processor i to processor j. There are no indexes, and all joins that are local to a processor are carried out using a sort-merge join. Assume that the relations are initially partitioned using a round-robin algorithm and that there are 10 processors. Describe the evaluation plan, briefly, to find the highest paid employee and give its cost in terms of t_d and t_s. You should also compute the 'elapsed time' cost (i.e., if several operations are carried out concurrently, the time taken is the maximum over these operations).	4
	(b)	Consider a relation 'r' over the attributes A, B, C with the following characteristics: • 5,000 tuples with 5 tuples per page • Attribute A is a candidate key • Unclustered hash index on attribute A (assume that the cost involve in searching index is 2 I/Os) • Clustered B+ tree index on attribute B • Attribute B has 1,000 distinct values in r • Attribute C has 500 distinct values and an unclustered 3-level B+ tree index (i) Estimate the cost of computing $\sigma_{A=constant}(r)$ using the index. (ii) Estimate the cost of computing $\sigma_{B=constant}(r)$ using the index.	6
9.	(a)	Write a MongoDB statement for each of the following SQL statements (Assume the name of database as 'mydb', name of collection as 'mycollection' and schema of R is (A, B, C, D,)):	4
	(b)	Write a Cassandra statement for each of the following: (i) Create a table (mytable) having - Attributes: A, B, C, D, & E (Assume suitable data type for each attributes). - Partition Keys: A & B - Cluster keys: D in ascending order and E in descending order. (ii) Write a CQL statement to retrieve all records of above created table whose attribute E value is 'p'.	6

Table for 9 (a):

(i) Select A, B, C From R Where D = 'p' or C = 'q';	(ii) Update R Set D = 'm' Where C = 'q' and A = 'n';
(iii) Select * From R Where A between 3 and 10;	(iv) Select A, C From R Where D in ('p', 'q', 'r');