

University of Dhaka
Department of Computer Science and Engineering
3rd Year 1st Semester Final Examination, 2022
CSE-3104: Database Management Systems - II (3 Credits)

Time: 3 hours

Total Marks: 70

Answer any five (5) out of the following seven (7) questions. Marks are given in the right margin.

- 1 (a) Define RAID. Explain the necessity of RAID with examples. [4]
- (b) What do you understand about P+Q Redundancy? Write notes on the RAID Level 5 with suitable examples. [5]
- (c) Explain different performance measures of disks. [5]
- 2 (a) Define indexing. Discuss the evaluation factors of an index structure. [5]
- (b) Make a B+ tree inserting the following values according to the given sequence. Consider the fanout value as 4. [Show only the tree after inserting each value]
23 → 11 → 67 → 34 → 97 → 13 → 42 → 53 → 7 → 1 → 100 → 99 → 90 → 3 → 77 [5]
- (c) "To make a primary index, one must choose the primary key as the search key" – explain with proper logic. [4]
- 3 (a) Explain the distinction between closed and open hashing. Discuss the relative merits of each technique in database applications. [3]
- (b) Suppose that we are using extendable hashing on a file that contains records with the following search-key values: 2, 3, 5, 7, 11, 17, 19, 23, 29, 31, 35, 44, 47, 49, 52, 59, 63. Show the extendable (dynamic) hash structure for this file if the hash function is $h(x) = x \bmod 8$ and buckets can hold three records. [6]
- (c) What are the causes of bucket overflow in a hash file organization? What can be done to reduce the occurrence of bucket overflows? [5]
- 4 (a) What do you understand by 'query processing' and 'query optimization'? Why is it not desirable to force users to make an explicit choice of a query processing strategy? [4]
- (b) Assume (for simplicity in this exercise) that only one tuple fits in a block and memory holds at most three blocks. Show the runs created on each pass of the sort-merge algorithm when applied to sort the following tuples on the first attribute: (kangaroo, 17), (wallaby, 21), (emu, 1), (wombat, 13), (platypus, 3), (lion, 8), (warthog, 4), (zebra, 11), (meerkat, 6), (hyena, 9), (hornbill, 2), (baboon, 12) [6]
- (c) Explain how you can apply the following equivalence rules to improve the efficiency of certain queries: [4]
- i) $E_1 \bowtie_{\theta} (E_2 - E_3) \equiv (E_1 \bowtie_{\theta} E_2 - E_1 \bowtie_{\theta} E_3)$
- ii) $\sigma_{\theta}(A \gamma_F(E)) \equiv A \gamma_F(\sigma_{\theta}(E))$, where θ uses only attributes from A
- 5 (a) Compare different partitioning techniques for I/O parallelism with respect to: [5]
- i) sequential search ii) point query iii) range query
- (b) Histograms are good for avoiding data distribution skew but are not very useful for avoiding execution skew. Explain why. How virtual node portioning technique is used to handle skew created by I/O partitioning? [2+3]
- (c) Explain the following forms of parallelism mentioning advantages and disadvantages: [4]
- i) Pipelined Parallelism
- ii) Independent Parallelism

- 6 (a) What do you understand by data analytics? What are the common steps in doing data analytics? [4]
- (b) How can you differentiate the following schemas used in data warehouse? Give examples. [4]
i) Star ii) Snowflake iii) Fact-constellation
- (c) The cube operation computes union of group by's on every subset of the specified attributes. [2]
Now consider the following cube operation on a multidimensional schema *sales* (*item_name*, *color*, *size*, *number*):

```
select item_name, color, size, sum(number)
from sales
group by cube(item_name, color, size)
```

Convert the above SQL using only rollup operation that will produce the same output.

- (d) Define the terms with respect the data mining: [4]
i) Classification ii) Regression
- 7 (a) What motivates you using object-based databases rather than relational database? [3]
- (b) Consider a schema for instructor. Each instructor has: [2+4]
- *ID*
 - *name* with sub-field *first_name* and *last_name* (composite attribute)
 - a list of *children* (multivalued attribute)
 - a list of *degree* achieved (multivalued attribute)
 - a set of *phone_nos* (multivalued attribute)
 - *age* (derived attribute of date attribute)
- i) Create 2 tuples for the nested relation based on above schema.
- ii) Show the 4NF decomposition of the nested relation mentioned in 7.b.i)
- (c) Suppose that you have been hired as a consultant to choose a database system for your client's application. For each of the following applications, state what type of database system (relational, persistent programming language-based OODB, object relational; do not specify a commercial product) you would recommend. Justify your recommendation. [5]
- i) A computer-aided design system for a manufacturer of airplanes.
 - ii) A system to track contributions made to candidates for public office.
 - iii) An information system to support the making of movies.


University of Dhaka
Department of Computer Science and Engineering
3rd Year 1st Semester Final Examination, 2021
CSE-3104: Database Management Systems II (3 Credits)

Time: 3 hours

Total Marks: 70

Answer any five (5) out of the following seven (7) questions. Marks are given in the right margin.

- 1 (a) What factors should be considered while choosing RAID levels? Compare RAID level 0, 1 and 5 with respect to these factors. [5]
- (b) Consider the deletion of record 5 from the file of Fig 1. Compare the relative merits of the following techniques for implementing the deletion: [3-3]



record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 11	98345	Kim	Elec. Eng.	80000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califhen	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000

Fig. 1

- i) Move record 6 to the space occupied by record 5, and move record 7 to the space occupied by record 6 and so on.
- ii) Move record 10 to the space occupied by record 5.
- iii) Mark record 5 as deleted, and move no records.

How free list can help to solve the problem? Explain with figures.

- (c) What is metadata? How it helps manipulating a database? [3]
- 2 (a) Consider the relation given below and the table X. Construct a B+ tree with $N = 4$, for indexing the table entries to perform the following query efficiently: [9]

Select * from X where Account Number = "AD0200001"

Account Number	Customer Name	Branch Name	Balance (Million)
AD0200019	ABC	C.DU	5
AD0200027	ABD	T.DU	7
AD0200001	ABB	N.K.D	7
AD0200021	ABE	RB.DU	9
AD0200025	ABBCD	C.DU	3
AD0200003	ABER	T.DU	5
AD0200009	YAJ	T.DU	6
AD0200008	KAL	RB.DU	77
AD0200007	PLL	C.DU	89
AD0200010	BPL	C.DU	90
AD0200018	IPL	C.DU	23
AD0200015	MPL	N.K.D	24
AD0200013	MMPL	RB.DU	54
AD0200030	XPL	N.K.D	57

- (b) Analyze the computational complexity of performing a query in a B+ tree with the node size N . Calculate the cost of performing the query stated in 2(a) with respect to the number of nodes accessed in a B+ tree during the query. [5]
- 3 (a) Let relations Patient (ID, Name, Address, Age) and Doctor (ID, Dept, Salary, Degree) have the following properties [7]
- The patient has 50000 tuples and needs 800 blocks in the secondary storage.
 - The doctor has 20000 tuples and needs 500 blocks in the secondary storage.

nested relation \rightarrow relation with
 no relation-valued attribute

A	B
a	(b,bb)

Perform Patient Doctor using the merge join algorithm when relations are not sorted according to ID (consider the cost of the last write operation for sorting).
 Compare the cost of the merge join algorithm with the nested and block nested-loop join algorithms.
 Assume: $M = 4$ blocks, $t_S = 5$ msec. and $t_T = 0.2$ msec.

- (b) With examples, show the differences among schedules, serial schedules, and serializable schedules. [3]
- (c) Consider the below-given schedule S. Proof and check whether it is conflict serializable or not? Also, check for the view serializability of the schedule. [4]

Transaction 1	Transaction 2
Read (A) $A := A - 50$ Write (A)	
	Read(B) $B := B - B * 0.1$ Write (B)
Read (B) $B := B + 50$ Write (B)	
	Read (A) $A := A + A * 0.1$ Write (A)

4 (a) 'Throughput' and 'response time' are two important performance measures of a database. How parallel systems help to improve these measures with respect to query processing? [2]

(b) Explain the factors that affect speedup and scaleup in parallel databases. [4]

(c) Compare 'shared memory' and 'shared disk' as parallel database architecture. [4]

(d) What are the benefits and limitations of using distributed database? [4]

5 (a) Define the terms: i) interquery parallelism ii) intraquery parallelism. [3]

(b) Describe the strategy for Range Partitioning Sort. How it differs from Parallel External Sort-Merge strategy? [5]

(c) When should we use Fragment-and-Replicate join? What is Asymmetric Fragment-and-Replicate join? [3]

(d) What is network partition in distributed databases? How it can be handled? [3]

6 (a) What are the purposes of using a data warehouse? Describe benefits and drawbacks of a source-driven architecture for gathering of data at a data warehouse, as compared to a destination-driven architecture. [4]

(b) What do you understand by OLAP? Distinguish between the OLAP operation:
 i) Slicing and dicing
 ii) cube and rollup [6]

(c) How do you define data mining? Briefly mention some of the applications of data mining. [4]

7 (a) What are the differences between object-relational database and object-oriented database? [3]

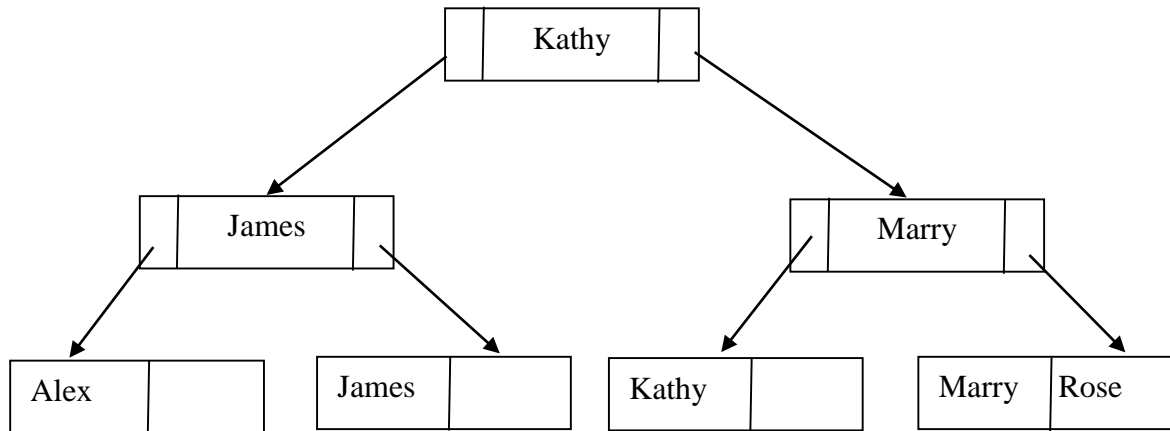
(b) What different complex data types are handled by the extensions to SQL that usual relational model does not support. Explain with a suitable example. [5]

(c) What do you understand by 'nesting' and 'unnesting'? [2]

(d) What is persistent programming language? How it differs from embedded SQL? [4]

b) Consider the following B+ - tree.

10



What would the tree look like after the following sequence of operations:
insert 'Bob', insert 'Marry', and delete 'Kathy'?

c) Consider a bucket split that occurs whenever an overflow page is created. Let $h_0(x)$ takes the rightmost 3 bits of key x as the hash value, and $h_1(x)$ takes the rightmost 2 bits of key x as the hash value.

10

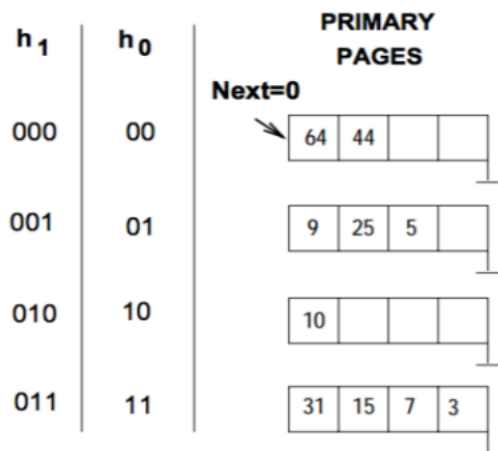


Figure-1: Hashing

Starting from the above hash table, plot the final hash table after inserting 17, 25, and 30. Remember to indicate the new hash function (if any), and to update the 'Next' pointer, if needed.

3. a) What are the steps involved in generating the query evaluation plan for an SQL expression? Briefly discuss with diagrams.

5

- b) 'Although indices are used for faster accesses to database, sometimes linear search could be a better choice' – explain. 4
- c) Find the query cost of block nested-loop join for worst and best case. Is indexed nested-loop join a better choice than block nested-loop join? Explain. 5
- d) Optimizers also use heuristics along with cost based optimization. How do heuristics help in this regard? 4.33
- e) What is the problem with 'materialization' approach of evaluation of an expression? If 'pipelining' approach is an alternative, what are the limitations of this approach? 5
4. a) State the pros and cons of data replication in distributed database systems. 3.33
- b) "Functional dependency (FD) has an impact on data fragmentation" – give your opinion with an explanation. 10

Consider the following 'Invoice' relation. Show horizontal and vertical fragmentations of this relation (note: you have to define its FDs before performing fragmentation).

Relation: INVOICE
Invoice_number
Product_Ino
Sale_date
Product_label
Product_price
Vendor_code
Product_quantity
Product_price

- c) Consider a distributed DBMS with four sites. In site 1, there is 'Employee' relation, in site 2, there is 'Department' relation and in site 3, there is 'Salary' relation. 10

In site 1, Employee:

ID	Name	DOB	Address	DeptNo
----	------	-----	---------	--------

Here, there are 200 records where each record is 100 bytes long. *ID* is 10 bytes long, *DeptNo* is 5 bytes long and *Name* is 10 bytes long.

In site 2, Department:

DeptNo	DeptName	Manager_ID
--------	----------	------------

Here, there are 10 records with each record is 30 bytes long. *DeptNo* is 5 bytes long, *DeptName* is 15 bytes long and *Manager_ID* is 10 bytes long.

In site 3, *Salary*:

DeptNo	ID	Salary
--------	----	--------

Here, there are 200 records with each record is 20 bytes long. *DeptNo* is 5 bytes long, *ID* is 10 bytes long and *Salary* is 5 bytes.

Suppose a user submits a query to site 4 in order to extract all manager names and their salaries. The result of this query will have 10 tuples if each department has one manager. Suppose each result tuple is 15 bytes long. Find an optimal way to execute this query so that the total data transfer be minimum.

5. a) Differentiate between 'Transaction Processing System' and 'Decision Support System'. 4
- b) With the help of an example define 'Support' and 'Confidence' of association rules. 4
- c) What are the motivations of using object based databases? 5
- d) Consider the following example of a non 1-NF relation *books*: 5.33

<i>title</i>	<i>author_array</i>	<i>publisher</i>	<i>keyword_set</i>
		(name, branch)	
Compilers	[Smith, Jones]	(McGraw-Hill, NewYork)	{parsing, analysis}
Networks	[Jones, Frick]	(Oxford, London)	{Internet, Web}

How are array [*author_array*] and multiset-valued [*keyword_set*] attribute declared in object based database? How can you create the table *books* using object-based approach?

- e) What are the advantages and drawbacks of persistent programming language? 5

University of Dhaka
Department of Computer Science and Engineering
3rd Year 1st Semester B. Sc. Final Examination 2019
CSE 3104: Database Management System II

Duration: 3 hours

Credits: 3

Full Marks: 60

(Answer any four of the following six questions)

1. a. What are the factors for choosing a RAID level? 3
 b. "RAID improves reliability via redundancy as well as performance via parallelism."- Explain properly. 4
 c. How variable-length records arise in database systems? Describe a technique to implement variable-length records while organizing file. 1+
4
 d. Indices speed query processing, but it is usually a bad idea to create indices on every attribute, and every combination of attributes, that is potential search keys. Explain why. 3
2. a. When a hash index is better than a B⁺-tree index? Explain with an example. 4
 b. Construct a dynamic hash index structure on search key Branch-name for the following database and hash addresses. Assume that each bucket can contain 2 entries 11

Hash function

Branch-name	h(Branch-name)
Brighton	0010
Downtown	1010
Mianus	1100
Perryridge	1111
Redwood	0011
Round Hill	1101

Database

Account-number	Branch-name	Balance
A-217	Brighton	750
A-101	Downtown	600
A-110	Downtown	600
A-215	Mianus	700
A-102	Perryridge	400
A-201	Perryridge	900
A-222	Redwood	700
A-305	Round Hill	350

3. a. Define the terms: i) Query processing ii) Query optimization. 3
 b. How can dynamic programming be used to handle a huge amount of evaluation plans in a query optimization process? 4
 c. Suppose that B⁺-tree index on *branch_city* is available on relation *branch* on schema *Branch(branch_name, branch_city, assets)*, and that no other index is available. List different ways to handle the following selections that involved negation: 3

i) $\sigma_{\neg (branch_city < "Brooklyn")}(branch)$

- ii) $\sigma_{\neg (branch_city = "Brooklyn")}(branch)$
- iii) $\sigma_{\neg (branch_city < "Brooklyn") \vee assets < 5000}(branch)$
- d. What do you understand by heuristics optimization? What steps are followed in typical heuristics optimization? 3
- e. Consider the materialized view $v = r \bowtie s$ and an update to r . Let set of tuples inserted to and deleted from r are denoted by i_r and d_r . Also r^{old} and r^{new} denote the old and new states of relation r . Now show the relational algebra expressions for incremental view maintenance. 2
4. a. What are the factors that limit speedup and scaleup in parallel database system? 2
- b. What different interconnection network architectures exist for parallel database? Mention characteristics of each with advantages and disadvantages. 4
- c. What factors could result in skew when a relation is partitioned on one of its attribute by: i) Hash partitioning ii) Range partitioning? In each case, what can be done to reduce the skew? 5
- d. Define the terms: 4
- i) Interquery and Intraquery parallelism
- ii) Pipelined and Independent parallelism
5. a. Differentiate homogeneous and heterogeneous distributed databases. 3
- b. What are the approaches of data fragmentation in distributed database? What advantage you may gain by these approaches? 4
- c. What is false cycle? How unnecessary rollbacks occur in global wait-for graph due to false cycle? 3
- d. Let r_1 be a relation with schema R_1 stores at site S_1 . Let r_2 be a relation with schema R_2 stores at site S_2 . Evaluate the expression $r_1 \bowtie r_2$ using semijoin strategy and obtain the result at S_1 . 3
- e. Consider a relation that is fragmented horizontally by *plant_number*: 2
- employee (name, address, salary, plant_number)*
- Assume that each fragment has two replicas: one stored at the New York site and one stored locally at the plant site. Describe a good processing strategy for the following queries entered at the San Jose site.
- i) Find all employees at the Boca plant.
- ii) Find the average salary of all employees.
- iii) Find the highest-paid employee at each of the following sites: Toronto, Edmonton, Vancouver, Montreal.
- iv) Find the highest-paid employee in the company.
6. a. What do you understand by a data warehouse? What are the design issues to be addressed in building a data warehouse? 2+
- b. Define star schema for a data warehouse with suitable example. 4
- c. What is persistent programming language? Mention the basic differences between object relational and object oriented data model. 3
- d. With suitable example explain unnesting and nesting of a nested relation. 1+
- e. 2
- f. 3