UNIVERSITY OF COLOMBO, SRI LANKA

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

DEGREE OF BACHELOR OF INFORMATION TECHNOLOGY (EXTERNAL)

Academic Year 2014/2015 –3rd Year Examination – Semester 6

IT6404 - Database Systems II

Structured Question Paper

19th July, 2015
(TWO HOURS)

To be completed by the candidate

BIT Examination Index No: ……………………………………

Important Instructions:

- The duration of the paper is **2 (two) hours**.
- The medium of instruction and questions is English.
- This paper has **4 questions** and **16 pages**.
- **Answer all questions (**25 marks each**)**.
- **Write your answers** in English using the space provided **in this question paper**.
- Do not tear off any part of this answer book.
- Under no circumstances may this book, used or unused, be removed from the Examination Hall by a candidate.
- Note that questions appear on both sides of the paper.
  If a page is not printed, please inform the supervisor immediately.

Questions Answered

Indicate by a cross (×), (e.g. ☒) the numbers of the questions answered.

	Question numbers				
To be completed by the candidate by marking a cross (×).	1	2	3	4	
To be completed by the examiners:					

1) (a) Why do databases use indexing techniques?

(02 marks)

ANSWER IN THIS BOX

Indexing is a data structure technique to efficiently retrieve records from the database files based on some attributes on which the indexing has been done.

[2]

(b) Briefly describe the following three indexing types.

(i) Primary Index

(ii) Secondary Index

(iii) Clustering Index

(06 marks)

ANSWER IN THIS BOX

(i) Primary index is defined on an ordered data file.

The data file is ordered on a key field.

The key field is generally the primary key of the relation. [2]

(ii) Secondary index may be generated from a field which is a candidate key and has a unique value in every record, or a non-key with duplicate values. [2]

(iii) Clustering index is defined on an ordered data file.

The data file is ordered on a non-key field.

[2]

(c) Ordered indexing is of two types, namely Dense index and Sparse index. Explain Dense and Sparse indexes identifying the differences between the two types.

(04 marks)

ANSWER IN THIS BOX

In dense index, there is an index record for every search key value in the database. This makes searching faster but requires more space to store index records itself. Index records contain search key value and a pointer to the actual record on the disk. [2]

In sparse index, index records are not created for every search key.

An index record here contains a search key and an actual pointer to the data on the disk.

To search a record, we first proceed by index record and reach at the actual location of the data.

If the data we are looking for is not where we directly reach by following the index, then the system starts sequential search until the desired data is found.

[2]

(d) Multi-level Index helps in breaking down the index into several smaller indices in order to make the outermost level so small that it can be saved in a single disk block, which can easily be accommodated anywhere in the main memory. Briefly explain the B+ Tree and its structure. Give a simple example of a B+ Tree structure to illustrate how the leaf nodes and internal nodes are connected.

(05 marks)

ANSWER IN THIS BOX

B+ tree is a balanced binary search tree that follows a multi-level index format. [1]

The leaf nodes of a B+ tree denote actual data pointers. [1]

B+ tree ensures that all leaf nodes remain at the same height, thus balanced.

Additionally, the leaf nodes are linked using a link list; therefore,

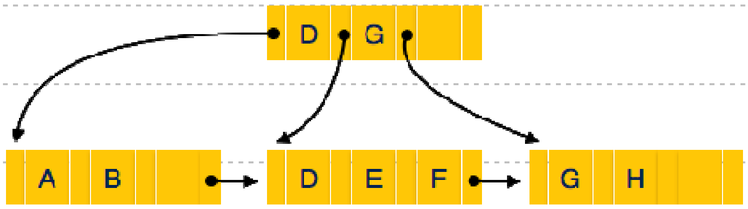
a B+ tree can support random access as well as sequential access. [1]

Continued…

Every leaf node is at equal distance from the root node.

A B+ tree is of the order n where n is fixed for every B+ tree.

e.g.



[2]

(e) Consider the following query issued on a supplier database consisting of several relations including the following three relations where primary keys are underlined and foreign keys are bold. Assume there are 200 suppliers located in 50 cities. Shipment table has 100,000 records with approximately 1,000 shipments per day involving all 200 suppliers equally.

```
supplier(sno, sname, city, status);
shipment(sno, pno, quantity, shipment_date);
part(pno, pname, description);

SELECT DISTINCT sname FROM supplier S,shipment sh
WHERE s.sno = sh.sno AND s.city = 'Colombo'
   AND sh.shipment_date = '01/06/2015';
```

(i) Assume that no indices are used for non-key fields. What is the most expensive way of processing the above query? Identify amount of records processed at the join and final result.

(04 marks)

ANSWER IN THIS BOX

Supplier and Shipment would be first joined (or Cartesian product) and thereafter the selection would be done. [1]

Due to primary keys sno of Supplier and be joined with corresponding Shipments. Hence 100,000 records would be processed before the selection. [1]

Of which only 1/50 will be from Colombo (i.e. 4) and 1/1000 would be for ‘01/06/2015’ [1]

As we want only unique supplier names, only 4 records would be the output. [1]

(ii) Suggest indexing techniques to make the above query efficient and explain how the query would be processed with the suggested indices.

(04 marks)

ANSWER IN THIS BOX

Indexing shipment_date and city as clustered index will help to identify those records. [1]

Selecting Suppliers from Colombo will give 200/50 = 4 records [1]

Selecting Shipments on '01/06/2015' will give 1000 = 1000 records [1]

Joining Supplier and Shipment for Colombo suppliers uniquely will give 4 = 4 records [1]

2) (a) (i) Consider the following schedule.

$w_1(a);\ r_2(a);\ w_1(b);\ w_3(c);\ r_2(c);\ r_4(b);\ w_2(d);\ w_4(e);\ r_5(d);\ w_5(e).$

Draw precedence graphs for the above schedule. Indicate if the graph has cycles or not. If the above schedule is a serializable schedule determine all the equivalent serial schedules and if not serializable, indicate why it is non-serializable and identify the type of conflict. Note that r_i and w_i denote respectively the read and write operations of transaction T_i for data item a, b, c, d & e.

(07 marks)

ANSWER IN THIS BOX

Serializable as it can swap all non-conflicting, e.g. $r_2(a); w_3(b);$ & $r_4(b); w_2(d);$ [1]

Three possible Serial Schedules: [3]

1. [T_1, T_3, T_2, T_4, T_5]

w1(a); w1(b); w3(c); r2(a); r2(c); w2(d); r4(b); w4(e); r5(d); w5(e)

Continued…

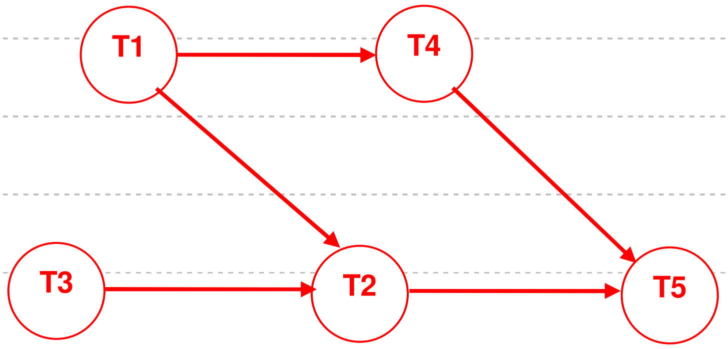
2. $[T_3, T_1, T_2, T_4, T_5]$

w3(c); w1(a); w1(b); r2(a); r2(c); w2(d); r4(b); w4(e); r5(d); w5(e)

3. $[T_3, T_1, T_4, T_2, T_5]$

w3(c); w1(a); w1(b); r4(b); w4(e); r2(a); r2(c); w2(d); r5(d); w5(e)

Graph has no cycles. [1/2]



[2 1/2]

[Candidate may give the diagram first and based on that decide if serializable.]

(ii) Assume that each transaction will commit at the earliest possible point of time soon after completing its last transactions in the schedule given in (i) above. If the schedule was executed under two phase locking protocol, write down the locks acquired, released or changed (i.e. 'Release S(a)' to indicate release of shared lock for a) including any waiting for locks, commits or deadlocks at each of the times starting at t1.

(08 marks)

ANSWER IN THIS BOX

Time	T1	T2	T3	T4	T5	Acquire Locks/Wait for	Release or Change Locks
t1	$w_1(a)$					X(a)	
t2		$r_2(a)$				Wait for S(a)	
t3	$w_1(b)$					X(b)	
t4	c_1						Release X(a) & X(b)
						T2 acquire S(a)	
t5			$w_3(c)$			X(c)	
t6			c_3				Release X(c)
t7		$r_2(c)$				S(c)	
t8				$r_4(b)$		S(b)	
t9		$w_2(d)$				X(d)	
t10		c_2					Release S(c) & X(d)
t11				$w_4(e)$		X(e)	
t12				c_4			Release S(b) & X(e)
t13					$r_5(d)$	S(d)	
t14					$w_5(e)$	X(e)	
t15					c_5		Release S(d) & X(e)

(b) Consider the following transaction log from the start of the run of a database system that is using undo/redo logging with checkpointing (CKPT) for crash recovery. The log entries for database updates are in the format: <Transaction id, Variable, New value, Old value>

```
1)    <START T1>
2)    <T1, A, 30, 10>
3)    <T1, B, 20, 0>
4)    <START T2>
5)    <T1, A, 60, 30>
6)    <T2, C, 10, 20>
7)    <COMMIT T1>
8)    <START T3>
9)    <T3, D, 50, 40>
10)   <T2, E, 40, 50>
11)   <CKPT (T2,T3)>
12)   <T2, C, 70, 10>
13)   <COMMIT T2>
14)   <START T4>
15)   <T4, F, 80, 70>
16)   <COMMIT T3>
17)   <T4, F, 100, 80>
18)   <COMMIT T4>
```

(i) Using the notations defined in (a) above, i.e. r_i and w_i produce the corresponding schedule for the above transaction log for all its entries.

(05 marks)

ANSWER IN THIS BOX

$W_1(A)$; [1/3]

$W_1(B)$; [1/3]

$W_1(A)$; [1/3]

$W_2(C)$; [1/3]

C_1; [1/2]

$W_3(D)$; [1/3]

$W_2(E)$; [1/3]

$W_2(C)$; [1/3]

C_2; [1/2]

$W_4(F)$; [1/3]

C_3; [1/2]

$W_4(F)$; [1/3]

C_4; [1/2]

(ii) What are the values of the data items A, B, C, D and E on disk after recovery if the system crashes just before line 11 is written to disk?

(03 marks)

ANSWER IN THIS BOX

A=60, B=20,

C=20, D=40,

E=50

(iii) What are the values of the data items A, B, C, D, E, and F on disk after recovery if the system crashes just before line 18 is written to disk?

(02 marks)

ANSWER IN THIS BOX

A=60, B=20,

C=70, D=50,

E=40, F=70

3) (a) Distribution transparency allows a physically dispersed database to be managed as though it were a centralized database. The level of transparency supported by the DDBMS varies from system to system. Three levels of distribution transparency are recognized, namely: Fragmentation transparency, Location transparency and Local mapping transparency. Briefly explain these three levels of distribution transparencies.

(06 marks)

ANSWER IN THIS BOX

Fragmentation transparency is the highest level of transparency.

The end user or programmer does not need to know that a database is partitioned.

Therefore, neither fragment names nor fragment locations are specified prior to data access. [2]

Location transparency exists when the end user or programmer must specify

Continued…

the database fragment names but does not need to specify where those fragments are located. [2]

Local mapping transparency exists when the end user or programmer must specify both the fragment names and their locations. [1]

(b) Consider the following Employee relation:

```
Employee(Name, DOB, Address, Department, Location, Salary)
```

Above Employee data are distributed over three different locations based on the location of the department, namely: Colombo, Galle and Kandy. The table is divided by location where Colombo employee data are stored in fragment E1, Galle employee data are stored in fragment E2 and Kandy employee data are stored in fragment E3.

(i) Now suppose the end user wants to list all employees with Salary greater than 50,000. Write the SQL statement to retrieve the above information under Fragmentation transparency.

(02 marks)

ANSWER IN THIS BOX

```
SELECT *
FROM Employee
WHERE Salary > 50000;
```

(ii) Now suppose the end user wants to list all employees with Salary greater than 50,000. Write the SQL statement to retrieve the above information under Location transparency.

(02 marks)

ANSWER IN THIS BOX

```
SELECT * FROM E1 WHERE Salary > 50000;
UNION
SELECT * FROM E2 WHERE Salary > 50000;
UNION
SELECT * FROM E3 WHERE Salary > 50000;
```

(iii) Now suppose the end user wants to list all employees with Salary greater than 50,000. Write the SQL statement to retrieve the above information under Local Mapping transparency.

(02 marks)

ANSWER IN THIS BOX

```
SELECT * FROM E1 NODE Colombo WHERE Salary > 50000;
UNION
SELECT * FROM E2 NODE Galle WHERE Salary > 50000;
UNION
SELECT * FROM E3 NODE Kandy WHERE Salary > 50000;
```

[Candidate may use appropriate keyword for NODE such as Site]

(c) The design of a distributed database introduces three new issues which are **not** present in a centralised database. Identify the three new issues and briefly explain each of them giving adequate details to identify how they are handled by a distributed database systems.

(09 marks)

ANSWER IN THIS BOX

The three issues are:

Continued…

How to partition the database into fragments?

Which fragments to replicate?

Where to locate those fragments and replicas? [3]

Data fragmentation allows to break an object into two or more fragments. [1]

The object might be a user's database, a system database, or a table.

Each fragment can be stored at any site over a computer network. [1]

Information about data fragmentation is stored in the distributed data catalogue, from which it is accessed by the Transaction Processors to process user requests.

Data replication refers to the storage of data copies at multiple sites served by a computer network. [1] Fragment copies can be stored at several sites to serve specific information requirements. Because the existence of fragment copies can enhance data availability and response time, data copies can help to reduce communication and total query costs. [1]

Data allocation describes the process of deciding where to locate data. [1]

Data allocation will be based on some allocation strategies. Data distribution over a computer network is achieved through data partition, through data replication, or through a combination of both. [1]

(d) What types of transparencies are provided to address the three issues in (c) above in a distributed database architecture? Briefly explain each of these transparencies.

(04 marks)

ANSWER IN THIS BOX

Fragmentation, Replication and Allocation transparencies. [1]

Fragmentation transparency: Data fragmentation is transparent to the user, who sees only one logical database. The user does not need to know the name of the database fragments in order to retrieve them. [1]

Replication transparency: The user sees only one logical database. The DDBMS transparently selects the database fragment to access. To the user, the DDBMS manages all fragments transparently. [1]

Allocation (Location) transparency: The user does not need to know the location of data in order to retrieve those data. [1]

4) (a) The star schema is a data modelling technique used to map multidimensional decision support data into a relational database. The basic star schema has four components: facts, dimensions, attributes, and attribute hierarchies. Briefly explain each of the four components.

(08 marks)

ANSWER IN THIS BOX

Facts are numeric measurements (values, e.g. Sales figures) that represent a specific business aspect or activity. [1]

Facts are normally stored in a fact table that is the center of the star schema. The fact table contains facts that are linked through their dimensions. [1]

Dimensions are qualifying characteristics that provide additional perspectives

Continued…

to a given fact. E.g. sales might be compared by product from region to region and from one time period to the next. [1]

Such dimensions are normally stored in dimension tables and linked to fact tables. [1]

Attributes are often used to search, filter, or classify facts.

Each dimension table contains attributes. [1]

Dimensions provide descriptive characteristics about the facts through their attributes. [1]

Attributes within dimensions can be ordered in a well-defined attribute hierarchy. [1]

The attribute hierarchy provides a top-down data organization that is used for two main purposes: aggregation and drill-down/roll-up data analysis. [1]

(b) Consider a multidimensional data model for students focussing on student attendance for classes conducted by lecturers. Assume there are three dimension table for this data model.

(i) What data could be selected to form the Fact table and what would be its dimension tables.

(02 marks)

ANSWER IN THIS BOX

Performance can be the Fact Table with data such as Attendance of relevant courses. [1]

Dimensions can be Student, Course, Lecturer, Location, Time [3 of them] [1]

(ii) Suggest possible attributes for Dimension Table proposed in (i) above.

(03 marks)

ANSWER IN THIS BOX

Student: Student No, Gender, Location;

Course: Name (code), Duration

Lecturer: Name, Gender

Location: Room

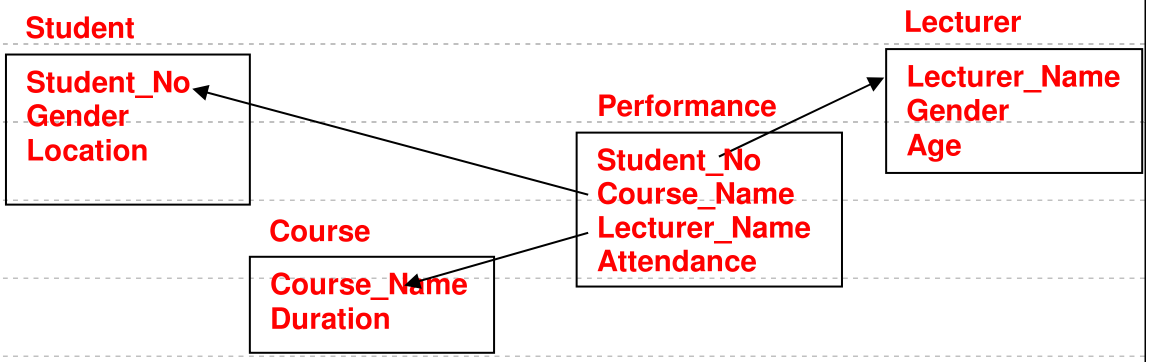
Time: Date, Time

Only 3 of them as identified in (ii) is required [3]

(iii) Draw a star schema for the above student data model.

(05 marks)

ANSWER IN THIS BOX



(iv) For the above Student model propose an attribute hierarchy for Location of students and state how it can be used to retrieve data.

(02 marks)

ANSWER IN THIS BOX

Attributes of Location namely Region, State, City, Store can form an attribute Hierarchy. [1]

It can be used to provide a top-down data organization that is used for aggregation and drill-down/roll-up data analysis. [1]

(c) Data mining process undergoes several phases to extract knowledge from data. Briefly explain the data preparation and classification/analysis phases and the type of activities performed at these phase giving examples where applicable.

(05 marks)

ANSWER IN THIS BOX

Data Preparation Phase [1/2]

This phase identify data sets from external and operation databases [1]

It also clean the data and integrate them. [1]

i.e. Gathering data, Describing, Exploring, Verifying quality,

Selecting data, Cleaning data, Constructing, Integrating, Formatting

Data analysis and classification phase [1/2]

Various types of data analysis is performed, such as [1]

Classification analysis, clustering and sequencing analysis, link analysis,

Trend and deviation analysis. [1]

****