UNIVERSITY OF COLOMBO, SRI LANKA

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

DEGREE OF BACHELOR OF INFORMATION TECHNOLOGY (EXTERNAL)

Academic Year 2010/2011 –3rd Year Examination – Semester 6

IT6403 - Database Systems II

Structured Question Paper

21st August, 2011
(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 **14 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) Define a transaction schedule.

(03 marks)

ANSWER IN THIS BOX
A transaction schedule (S) is a list of operations showing the order of execution of a set of transactions (T1, T2, ..., Tn) over time.
The operations of Ti in S must appear in the same order in which they occur in Ti.

(b) Consider the following two transactions T1and T2. The value of "A" before executing any of the two transactions is 100.

T1
Read(A)
A = A + 100
Write(A)
B = A +50
Write(B)
Commit

T2
Read(A)
A = A + A*10%
Write(A)
Commit

(i) Give a serial schedule for the above two transactions.

(04 marks)

ANSWER IN THIS BOX
T1: Read(A)
T1: A = A + 100
T1: Write(A)
T1: Write(A)
T1: B = A + 50
T1: Write(B)
T1: Commit
T2: Read(A)
T2: A = A + A*10%
T2: Commit
OR Operations of T2 followed by operations of T1.

(ii) If the above transactions were executed as a serializable schedule, what could be the possible values for "A"?

(03 marks)

ANSWER IN THIS BOX

220 or 210

(iii) Give a conflict serializable schedule for the above two transactions.

(05 marks)

ANSWER IN THIS BOX

One possible non-serial schedule is given below.

T1: Read(A)

T1: A = A + 100

T1: Write(A)

T2: Read(A)

T2: A = A + A*10%

T2: Write(A)

T2: Commit

T1: B = A + 50

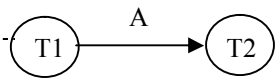
T1: Write(B)

T1: Commit

(iv) Construct the precedence graph for the schedule in (iii) above.

(03 marks)

ANSWER IN THIS BOX



(c) Compare the two phase locking protocol with the timestamp protocol identifying advantages and disadvantages of the two protocols.

(07 marks)

ANSWER IN THIS BOX

Two phase protocol uses lock while timestamp does not.

Due to locks, two phase can come across deadlocks while timestamp does not.

However deadlock prevention protocol of two phase can be used to ensure that the transaction will never enter a deadlock state.

Here, each transaction locks all its data before it begins execution.

Hence either all requested data items are locked in one step or none is locked.

Disadvantages are, low data utilisation where many data items may be locked but data is unused for a long period of time and

possible starvation where a transaction which requires a number of data items for its operation may find itself in a indefinite wait state while at least one of the data items required is always locked by some other transaction. To overcome this, deadlock detection method can be used.

Timestamp ordering is another method to determine serializability.

Here, there is no deadlock and no locks.

Basic idea is, if a transaction A starts before transaction B, then A should behave as if it is completed in entirety before B is started.

i.e. as a serial schedule.

2) (a) Discuss the difference between the "differed update" and "immediate update" particularly with respect to use of the log files for recovery purposes.

(03 marks)

ANSWER IN THIS BOX

Differed update changes the database only at the commit point and hence the log entries are not usually required for recovery purposes.

Immediate update changes the database before the commit point and hence the log file is required to recovery purposes.

(b) Architecture for distributed databases is different from that of a centralised database. List the type of differences that one could observe in a Distributed Data Base Management System (DBMS) when compared with a centralised DBMS.

(04 marks)

ANSWER IN THIS BOX

Distributed databases are under the control of a central DBMS in which storage devices are not all attached to a common CPU.

In Distributed DBMS, each site is a database system site in its own right, but the sites have agreed to work together (if necessary).

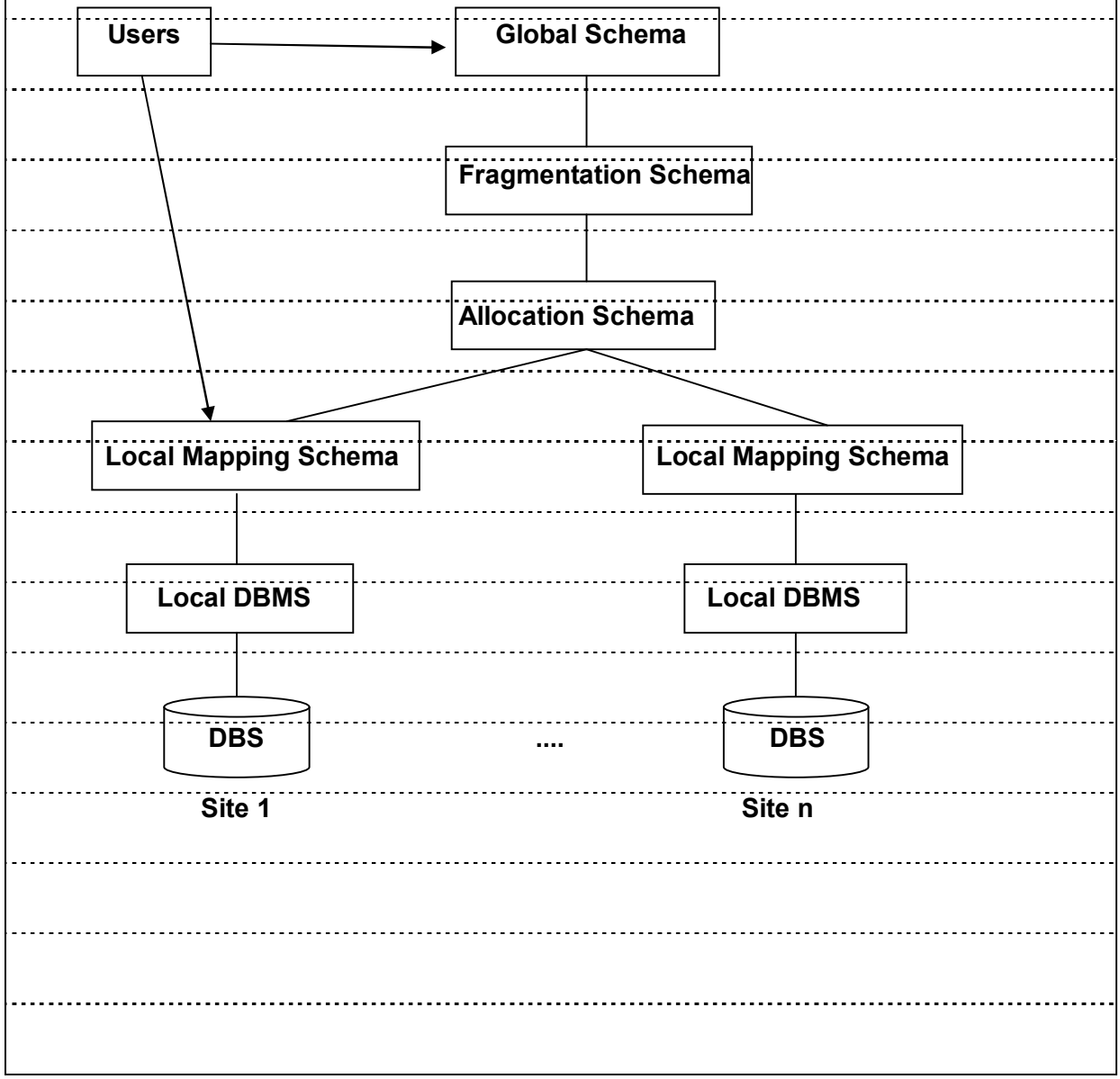
User at any site can access data anywhere in the network exactly as if the data were all stored at user's own site. It is a centralised view with distributed data.

A node in a distributed database system acts as a client, a server or both, depending on the situation.

(c) Draw the referenced architecture for a Distributed Data Base Management System identifying its components.

(06 marks)

ANSWER IN THIS BOX



(d) Consider the following global relations of a distributed database system having two departments.

```
Employee(EmpNo, EmpName, Designation, dob, BasicPay, HoursWorked, DeptNo)
Department(DeptNo, DeptName, Head, Phone)
```

If the following three queries represent the most frequently used local queries of a distributed DBMS issued at three different sites, namely Finance, Department 1 and Department 2. Sites have agreed to share data so that a global user could access the above Employee and Department relations.

```
(I)   SELECT EmpNo, BasicPay, HoursWorked FROM Salary;
(II)  SELECT EmpNo, EmpName, Designation, dob, DeptNo FROM Emp
             WHERE DeptNo="1";
(III) SELECT DeptNo, DeptName, Head, Phone FROM Dept WHERE DeptNo="2";
```

Discuss the fragmentation schema that could be used for the above data. Identify the fragmentation type for each relation and the construction process of the global relations.

(08 marks)

ANSWER IN THIS BOX

Query (II) indicates that Emp is required to be fragmented Horizontally by DeptNo attribute.

The two Emp fragments of the two departments must be merged using the UNION operator to give an intermediate relation named E1.

To view data of the Employee relation, Salary relation of query (I) and relation E1 have to be joined over EmpNo attribute.

Thus the Salary relation is a Vertical fragment of the Employee relation and Emp relation is a mix fragment of the Employee relation.

Query (III) indicates that the Dept relation is fragmented Horizontally by DeptNo attribute.

Thus Dept relation fragments belonging to each department must be merged using the UNION operator giving the relation Department.

(e) An empty Student database is to be populated with data. Following steps indicate the operations that took place with respect to the Student database.

Step	Date	Time	Operation
1	01-Jul-2011	09.00	Add two Student records
2	01-Jul-2011	09.30	Modify one Student record
3	01-Jul-2011	10.30	Delete one Student record
4	02-Jul-2011	09.00	Add one Student record

Assuming that the database is automatically backed up at midnight each day, indicate the number of student records in the database and the total number log entries after each of the above steps.

(04 marks)

ANSWER IN THIS BOX

Step	Number of Student Records	Total number of Log Entries
1	**2**	**2**
2	**2**	**3**
3	**1**	**4**
4	**2**	**1**

3) (a) Consider the following schema for a portion of a simple company database and the two SQL statements executed using two different query plans.

```
Employee(EmpNo, Name, Job, dob, Pay)

(I) SELECT Name, Pay FROM Employee WHERE Designation="Manager";

(II) SELECT Name, Job FROM Employee WHERE Pay>"3000" and Pay<"5000";
```

(i) Suggest the most suitable index file organisation for the query (I) above and give reasons.

(03 marks)

ANSWER IN THIS BOX

Harsh index is most suitable for querying individual records.

Hashed file organisation is a storage structure in which the address for each record is determined using a hashing algorithm.

Hashing algorithm is a routine that converts a key value into a relative record number.

This record number is to be used to determine storage location of the record.

Same is used later to directly access individual records using its key value.

(ii) Suggest the most suitable index file organisation for the query (II) above and give reasons for your choice.

(04 marks)

ANSWER IN THIS BOX

B+ tree is the most suitable index file for range level access.

A B+ tree is a type of tree which represents sorted data. It allows for efficient insertion, retrieval and removal of records, each of which is identified by a key.

It is a dynamic, multilevel index, with maximum and minimum bounds on the number of keys in each index segment.

In a B+ tree, all records are stored at the leaf level of the tree.

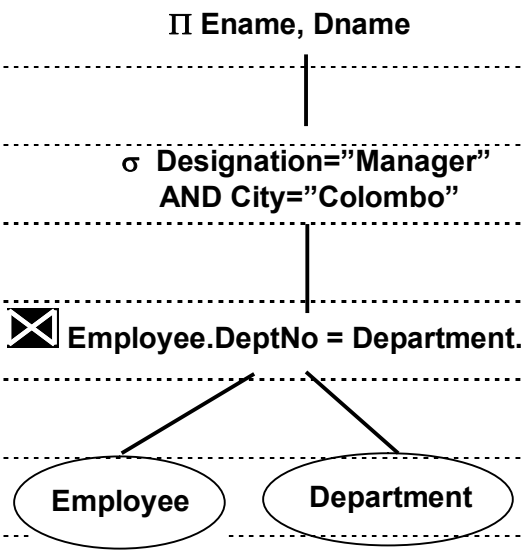
(b) Consider the following SQL query.

```
SELECT E.Ename, D.Dname FROM Employee E, Department D WHERE
E.Designation='Manager' AND E.DeptNo=D.DeptNo AND D.City='Colombo';
```

(i) The above query is to be processed by the query optimiser. Map the query into an initial query tree representation.

(03 marks)

ANSWER IN THIS BOX



(ii) Write Relational Algebra expression(s) for the initial query tree given in (b) (i) above.

(04 marks)

ANSWER IN THIS BOX

Join Employee and Department over DeptNo giving R1

Restrict R1 where Designation="Manager" and City="Colombo" giving R2

Project R2 over Ename and Dname giving Result

OR

$\Pi_{\text{Ename, Dname}} (\sigma_{\text{Designation="Manager" AND City="Colombo"}} (E \infty_{\text{D.deptNo=E.DeptNo}} D))$

(iii) What transformation rules should be used to optimise the query given in (b)(ii)? For each transformation rule show the transformation process using (a) relational algebra expression(s).

(06 marks)

ANSWER IN THIS BOX

cascade of σ

$\Pi_{\text{Ename, Dname}} (\sigma_{\text{Designation="Manager" AND City="Colombo"}} (E \infty D)) \rightarrow$

$\Pi_{\text{Ename, Dname}} (\sigma_{\text{Designation="Manager"}} E \infty_{\text{D.deptNo=E.DeptNo}} \sigma_{\text{City="Colombo"}} D)$

commutative of σ

$\Rightarrow \Pi_{\text{Ename, Dname}} (\sigma_{\text{City="Colombo"}} D \infty_{\text{D.deptNo=E.DeptNo}} \sigma_{\text{Designation="Manager"}} E)$

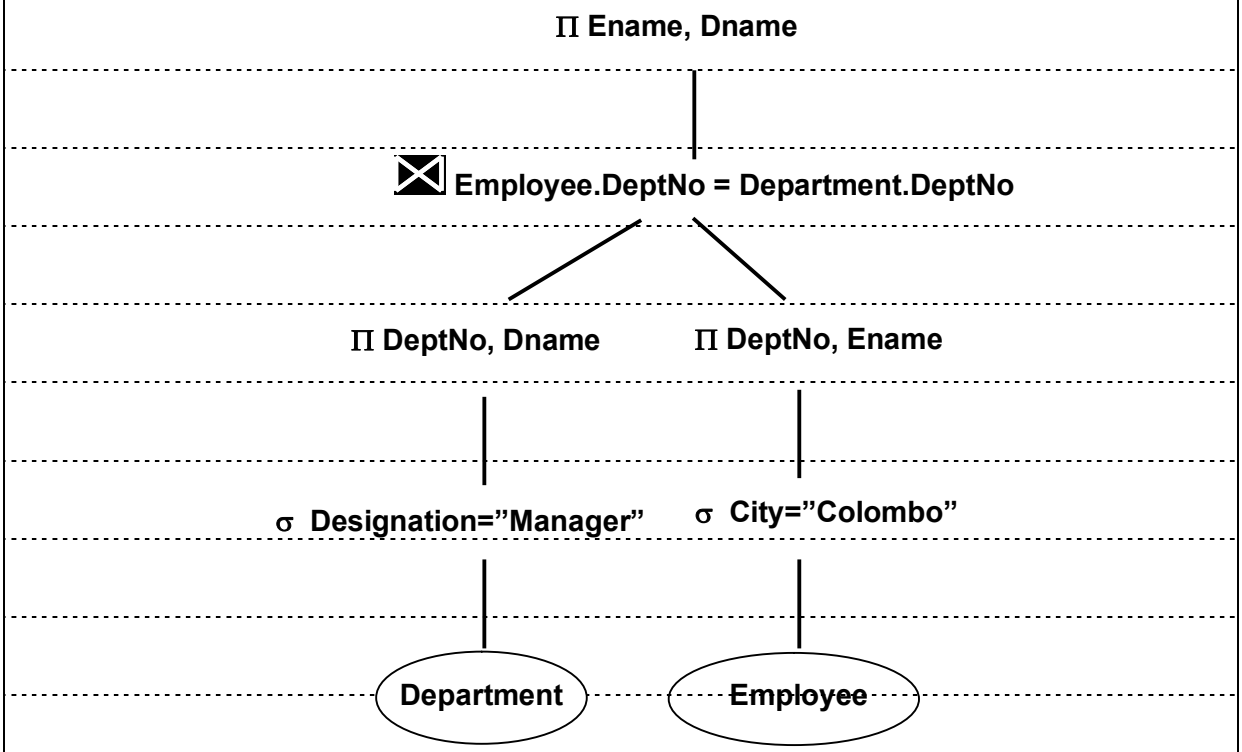
commutative of σ with π

$\Rightarrow \Pi_{\text{Ename, Dname}} (\Pi_{\text{Ename, Dname}} (\sigma_{\text{City="Colombo"}} D) \infty_{\text{D.deptNo=E.DeptNo}} \Pi_{\text{Ename, Dname}} (\sigma_{\text{Designation="Manager"}} E))$

(iv) Draw the optimized query tree for the query in (b).

(05 marks)

ANSWER IN THIS BOX



4) (a) Type constructor *row* is used to specify complex types known as user-defined types. Give an example to specify a row type for an address that would include street, city and zip. Also show how this row type address can be used to create a customer table having customer-id, customer-name, customer-address and phone.

(05 marks)

ANSWER IN THIS BOX

```
CREATE ROW TYPE Address (

  Street VARCHAR(25),

  City    VARCHAR(20),

  Zip     VARCHAR(9) );


CREATE TABLE Customer (

Customer-ID            CHAR(10),
```

Continued...

Customer-Name VARCHAR(25),

Customer-Address Address,

Phone VARCHAR(15));

(b) Show how Customer 'A.B. De Silva' with ID '12345', Street '67, Galle Road', City 'Colombo', Zip '00400' and Phone '0112344344' can be inserted as a record into the Customer table created in (a) above.

(03 marks)

ANSWER IN THIS BOX

```
INSERT INTO Customer (Customer-ID, Customer-Name, Customer-Address, Phone)
VALUES('12345', 'A.B. De Silva', ('67, Galle Road', 'Colombo', '00400'),
       '0112344344');
```

(c) In the table given below, **Column I** contains typical functions of a Data Warehouse. **Column II** lists description of some data warehouse functionality.

	Column I		**Column II**
A	Roll-up	1	Data is available by value or range.
B	Drill-down	2	Performing projection operations on the dimensions
C	Pivot	3	Cross tabulation is performed.
D	Slice and dice	4	Data is summarised with increasing generalization.
E	Selection	5	Increasing levels of details are revealed.
		6	Attributed are computed by operations on stored values.
		7	Data is sorted by ordinal value.

Match each function from **Column I** with the most appropriate description in **Column II**. Write your answer in the box given below the table.

(05 marks)

ANSWER IN THIS BOX

A – **4** B – **5**

C – **3** D – **2**

E – **1**

(d) During the data mining process, different types of knowledge are discovered. Using an example, briefly describe the knowledge that is expected to be discovered through Association rules and Sequential patterns.

(05 marks)

ANSWER IN THIS BOX

Association rules discover interesting relationships between variables.

E.g. If a person buys tea how likely is it for him to buy sugar.

Sequential patterns discover events with temporal relationships.

E.g. If a person buys a computer and later CD ROMs, then how likely is it for him to by a webcam a few months later.

(e) A market basket analysis is to be performed to find the relationships between a set of items = {milk, tea, butter, sugar}. After inspecting five baskets for these items the following were found.

```
B1 = {milk, tea, sugar}          B2 = {tea, butter, sugar}
B3 = {butter, sugar}             B4 = {milk, tea}
B5 = {tea, sugar}
```

(i) Determine the support for the following two rules: `tea => milk` and `tea => sugar`.

(04 marks)

ANSWER IN THIS BOX

Tea and milk appear in 2 out of 5 baskets giving a support of 40%.

Tea and sugar appear in 3 out of 5 baskets giving a support of 60%.

(ii) What is the confidence of tea, given milk and sugar (i.e. association rule {milk, sugar} => tea)?

(03 marks)

ANSWER IN THIS BOX

Milk and sugar appear together in only 1 basket.

Tea also appears in this basket.

Thus the confidence of tea given milk and sugar is = 100%.

****