UNIVERSITY OF COLOMBO, SRI LANKA

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

DEGREE OF BACHELOR OF INFORMATION TECHNOLOGY (EXTERNAL)

Academic Year 2012/2013 –3rd Year Examination – Semester 6

IT6404 - Database Systems II

Structured Question Paper

3rd August, 2013

(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) (i) Define Structured, Semi-Structured and Unstructured Data, with examples.

(09 marks)

ANSWER IN THIS BOX

Structured Data

A relation in a relational database contains several records and each record has a format consistent with other records in that relation. Such type of data that is represented in a strict format is called structured data.

(data is organised in semantic entities; similar entities are grouped together; entities in the same group have the same attributes; attributes have the same defined format with a predefined length and are all present in the same order)

E.g. Book will have a particular structure (e.g. table) to record title, authors, publisher, edition, year etc.

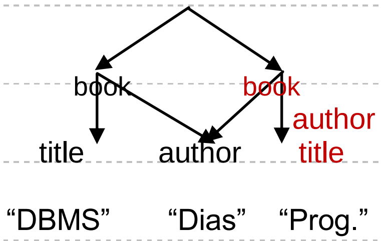
Semi-Structured Data

In some applications, data is collected in an ad hoc manner before having information regarding how it will be stored and managed. Although the collected data may have certain structure, but not all the data have an identical structure. Different entities may have different sets of attributes that means there is no pre-defined schema. Such type of data is known as semi-structured data.

(organised in semantic entities; similar entities are grouped together; entities in same group may not have same attributes; order of attributes not necessarily important; not all attributes may be required; size of same attributes in a group may differ; type of same attributes in a group may differ)

E.g. Data model for Book representing relationships between book, author, title etc., resulting in a graph structure.

(diagram optional)



Continued…

Unstructured Data

The term unstructured refers to the fact that no identifiable structure within this kind of data is available. Unstructured data cannot be stored in rows and columns in a relational database.

(data can be of any type; not necessarily following any format or sequence; does not follow any rules; is not predictable; examples include)

An example for unstructured data is documents/articles/chapters that are archived in a file folder. Other examples are videos and images.

(ii) Which structures and formats can be used to record above defined types of Data?

(05 marks)

ANSWER IN THIS BOX

Structured - Relational database tables

Table structures

Semi-Structured - XML/RDF format

Tree structures

Unstructured - Character and binary data

(b) What representational issues should be considered when converting data from a database into XML?

(02 marks)

ANSWER IN THIS BOX

Create the appropriate XML hierarchy and corresponding XML schema document.

(c) (i) "A data warehouse can be generally described as a technology aimed at enabling the knowledge worker to make better and faster decisions". Justify this statement.

(03 marks)

ANSWER IN THIS BOX

Data warehouse includes large volumes of historical data from multiple data sources that can be aggregated and disaggregated to extra knowledge for decision making workers such as a CEO, Directors and Managers.

(ii) A data warehouse uses a multidimensional storage model involving two types of tables. Name and describe how they are used.

(06 marks)

ANSWER IN THIS BOX

Dimension tables and fact tables

A dimension table consists of tuples of attributes of the dimension.

A fact table can be thought of as having tuples, one per each recorded fact.

This fact contains some measured or observed variable(s) and identifies it (them) with pointers to dimension tables. The fact table contains the data (e.g. Amount and Quantity Sold for a Product during a period), and the dimensions identify each tuple in that data (e.g. Product will identify Product No, Name, Description).

2) (a) Distribution of data and programs across the sites of a computer network is carried out during the design of a distributed database system. What is a typical reasonable unit of distribution for such design? Justify your answer.

(02 marks)

ANSWER IN THIS BOX

A fragment of a relation is a reasonable unit of distribution unlike the whole relation.

Most application views are usually subsets of relations (fragments) than the whole relation.

(b) (i) What are the advantages of fragmentation?

(03 marks)

ANSWER IN THIS BOX

improve reliability, performance, balance storage capacity and costs, communication costs, security

Locality of accesses of applications is defined on subsets of relations.

Execute concurrent transactions by accessing different portions of a relation.

Parallel execution of a single query (intra-query concurrency).

Reduce the volume of remote data accesses.

(ii) What is the qualitative and quantitative information that can be used to decide fragmentation?

(02 marks)

ANSWER IN THIS BOX

Quantitative information: frequency of queries, site, where query is run, selectivity of the queries, etc.

Qualitative information: types of access of data, read/write, etc.

(c) Consider the following project relations.

Project (<u>ProjNo</u>, pName, Budget, Location)

Following rules applies to the above relation.

- Project budget is to be hidden from most employees and hence should be separated from other attributes.
- Projects with a budget less than Rs. 200,000 should be separated from the rest as higher management approval is required for those costing Rs. 200,000 or above.
- Project monitoring process for those located outside Colombo is different to those in Colombo and hence managed separately.

(i) If a distributed database is to be designed, based on the above rules, give a set of predicates to partition the Project relation. Identify the resultant relations and the type of fragmentation applied.

(10 marks)

<u>ANSWER IN THIS BOX</u>

Project is vertically fragmented as Proj1 and Proj2 to separate Budget from rest.

Proj1 = $\Pi_{ProjNo,Budget}$ (Project)

Proj2 = $\Pi_{ProjNo,pName,Location}$ (Project) [3]

Proj1 is fragmented horizontally as LowBudProj and HighBudProj to separate low cost projects (< 200000) from the rest. Resultant relations are mixed fragmented.

LowBudProj = σBudget<200000 (Proj1)

HighBudProj = σBudget≥200000 (Proj1) [3]

Proj2 is further fragmented horizontally as ColProj and OtherProj to separate Colombo projects from the rest. Resultant relations are mixed fragmented.

ColProj = σLocation=”Colombo” (Proj2)

OtherProj = σLocation<>”Colombo” (Proj2) [3]

LowBudProj(<u>ProjNo</u>, Budget)

HighBudProj(<u>ProjNo</u>, Budget)

Continued…

ColProj(<u>ProjNo</u>, pName, Location)

OtherProj(<u>ProjNo</u>, pName, Location) [1]

(ii) Explain the correctness rules of fragmentation. Apply them to c(i) above.

(08 marks)

ANSWER IN THIS BOX

Completeness - Decomposition of relation R into fragments R_1, R_2, . . . , R_n is complete iff each data item in R can also be found in some R_i.

Reconstruction - If relation R is decomposed into fragments R_1, R_2, . . . , R_n, then using union or join relational operators appropriately, R should be reconstructed from its fragments.

Disjointness - If relation R is decomposed into fragments R_1, R_2, . . . , R_n and data item d_i appears in fragment R_j , then d_i should not appear in any other fragment R_k, $k \neq j$ (exception: primary key attribute for vertical fragmentation)

LowBudProj U HighBudProj → Proj1

ColProj U OtherProj → Proj2

Proj1 ⋈ Proj2 → Project

3) (a) Consider the schedules given in S1 and S2 below. Please note that r_i and w_i denote respectively the read and write operations of transaction T_i. a, b, c are data items.

```
S1 =  r₁(a), r₂(c),r₁(c), r₃(a),r₃(b), w₁(a),w₃(b),r₂(b),w₂(c),w₂(b)
S2 =  r₁(a), r₂(c),r₃(a), r₁(c),r₂(b), r₃(b),w₁(a) ,w₂(c),w₃(b),w₂(b)
```

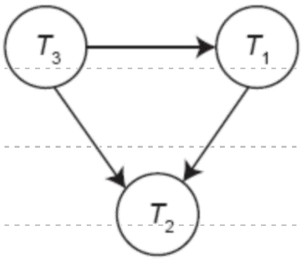
Produce the precedence graphs and determine whether the given schedules S1 and S2 are serializable. If so give the corresponding serial schedule.

(06 marks)

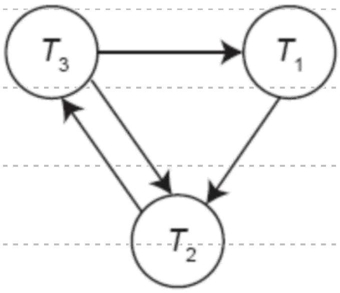
ANSWER IN THIS BOX

Based on the precedence graph, S1 is serializable and the serialization order is

T3 → T1 → T2 .



Based on the precedence graph, S2 is not serializable as there is a cycle in the graph.



(b) Explain what is meant by a recoverable schedule and why recoverability of schedules is desirable.

(04 marks)

ANSWER IN THIS BOX

A recoverable schedule is one where, for each pair of transactions Ti and Tj such that Tj reads data items previously written by Ti, the commit operation of Ti appears before the commit operation of Tj.

Continued…

Recoverable schedules are desirable because failure of a transaction might bring the system into an irreversibly inconsistent state.

(c) Consider the following schedule.

$r_3(c)$, $w_3(c)$, $r_1(a)$, $r_2(b)$, $w_2(b)$, $w_1(a)$, $r_1(b)$, $w_3(a)$, C1, C2, C3;

State whether the given schedule is recoverable or not. If the schedule is not recoverable propose the necessary modifications to make it recoverable.

(04 marks)

ANSWER IN THIS BOX

The given schedule is not recoverable, since T1 reads a value of b updated by T2, and yet, T1 commits before T2.

r3(c), w3(c), r1(a), r2(b), w2(b),w1(a),r1(b),w3(a), C2, C1, C3;

(d) Consider the following two schedules.

$S1 = r_1(a)$, $r_2(b)$, $w_2(b)$, $r_1(b)$;

$S2 = w_1(a)$, $w_2(a)$, $w_3(a)$, $w_1(b)$, $w_2(b)$;

Based on the two schedules S1 and S2 given below justify giving reasons which schedule is allowed by locks and not by timestamps and which schedule is allowed by timestamps and not by locks. Please note that r_i and w_i denote the read write operations for data items a and b of transaction T_i.

(06 marks)

ANSWER IN THIS BOX

S1 is allowed by locks and not by timestamps since *read X* is allowed only if *t* ≥ w-max(X). Since t1-read(B) < w-max(B), this condition is violated and as such the timestamp protocol cannot be adopted.

Continued…

However, it is possible to acquire locks based on 2PL protocol (as T1 may wait in its growing phase till T2 releases locks on B).

S2 is allowed by timestamps and not by locks since it is not possible to acquire locks based on 2PL protocol. Hence S2 is a serializable schedule which is illegal under the 2PL protocol.

(e) Consider the following part of a log file.

```
Start(T1)
Write(T1, x, old, new)
Commit(T1)
checkpoint
Start(T4)
Write(T4, y, old, new)
Write(T4, z, old, new)
Commit(T4)
Start(T2)
Write(T2, y, old, new)
Start(T3)
Write(T3, z, old, new)
    system crash
```

Explain giving reasons, which transactions could be recovered and which transactions cannot be recovered from the above log file based on Deferred Update mechanism.

(05 marks)

<u>ANSWER IN THIS BOX</u>

Since T1 and T4 are committed, their changes were written to disk.

However, T2 and T3 did not commit, hence their changes were not written to disk.

In the recovery process, the transactions that did not commit would be ignored.

4) (a) Consider the following relational schema where Id is the primary key in Student.

```
Student(Id, Name, Stream)
Enroll(StudId, Course)
```

Consider the query

```
SELECT * FROM Student S, Enroll E
WHERE S.Id = E.StudId AND E.Course = 'CS305'
                    AND S.Stream= 'Networking';
```

Suggest the indexes that can be added to this database in order to improve the performance of this query. Indicate whether these indices should be clustered or not. Explain your reasoning briefly.

(05 marks)

ANSWER IN THIS BOX

Clustered hash index on Course in Enroll can improve the performance, since it will facilitate the first selection (i.e. course='CS305').

A clustered index on Stream in Student might also help, since it can facilitate the second selection (i.e. stream='Networking').

Note that since Id is a primary key in Student, it might have an unclustered (rather than clustered) index, because numeric single-attribute keys often have unclustered hash indices.

We could do away with the clustered index on Stream, since we could also do the join using the index on Id, which exists because Id is the primary key.

(b) The table Employee(Id, Name, DeptId, Salary) has Id as its primary key and consider the following SQL statement issued on the Employee relation.

```
SELECT E.Name
FROM Employee E
WHERE E.Salary > 100000 AND E.DeptName = 'accounting';
```

What index would you choose to enhance the performance of the SQL statement given above considering each of the following scenarios separately?

(i) The number of employees in accounting is smaller than the number earning over Rs.100000.

(ii) The number of employees in accounting is much larger than the number earning over Rs. 100000.

(08 marks)

ANSWER IN THIS BOX

An index on either Salary or DeptName is needed depending on which is more selective.

If the number of employees in accounting is smaller than the number earning over Rs.100,000 then a hash or B+ tree index on DeptName would be appropriate.

Performance can be further improved by making the index clustered, so that all employees in the same department are in the same hash bucket or are consecutive if a B+ structure is used.

In that case, the index on Id must by unclustered, but that should not imply a performance penalty since queries involving Id will generally not involve a range search and hence will return a single row.

If accounting is a large department, then a B+ tree index on Salary would be appropriate since a range search is needed.

The same considerations with respect to clustering apply.

(c) Consider the following relations that represent part of a company database:

```
Employee(EmpId, Ename, Designation, Salary)
Works_On(EmpId, ProjId, Hours)
```

The Employee relation keeps information on employees and the Works_On relation has information on who is working on which project. Key of each relation is underlined. An employee may work for many projects and vice versa. However, there are Employees who may not work for any of the projects. There may be about 200 employees per project and only 10% of them are working more than 20 hours for a project. There are about 3,000 employees working for the company and 50% of them are earning more that Rs. 40,000.

Consider the following query:

```
   SELECT E.Ename, E,Designation
   FROM  Employee E, Works_On W
   WHERE E.Empid=W.Empid AND W.ProjId = '007'
                         AND W.Hours > 20  AND  E.Salary > 40000;
```

Assume that the following statistics and indices are available.

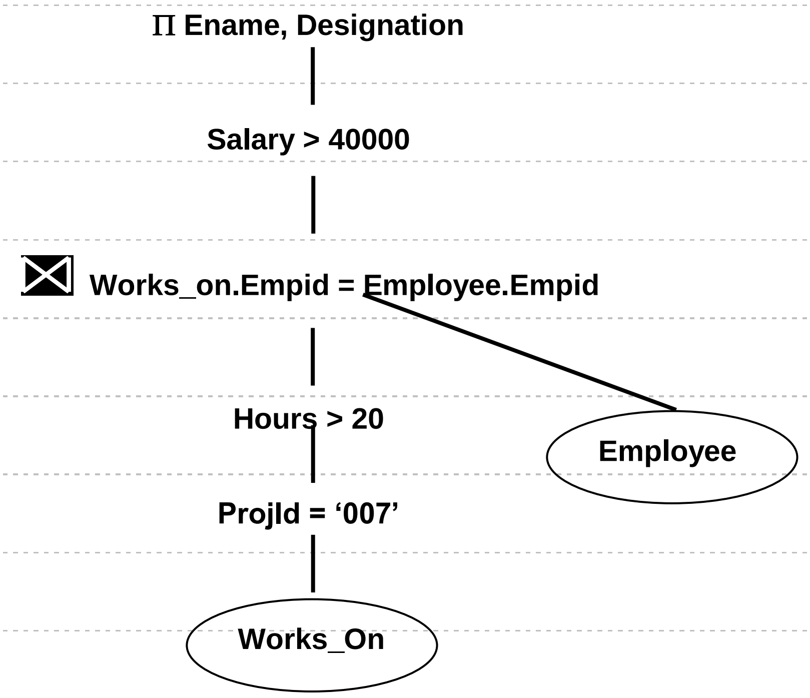
Works_On: 10,000 records on 50 projects, 25 records/page
Clustered static hash index on ProjId

Employee: 3,000 with 10 tuples/page
Unclustered hash index on EmpId

(i) Draw the query tree corresponding to the best query plan of the above query.

(06 marks)

ANSWER IN THIS BOX



(ii) Estimate the cost of the best query plan using the I/O cost. Note that the estimated cost to retrieve an unclustered page is about (1.2) I/O on average for hash based indexes.

(06 marks)

ANSWER IN THIS BOX

The select ProjId='007' from works_on would produce 200 tuples (i.e. 8 pages) and

hence costs 8 I/O due to clustered hash index.

Continued…

Selection of Hours > 20 is applied on the fly and that will reduce the number of tuples to 20 (i.e. 200 * 10%) which is 1 page.

The cost of finding matching Employee tuple is (1.2) I/O based on hash indexing.

As there are 20 such tuples the cost for retrieving the corresponding Employee tuples is 20 * (1.2) I/O => 24 I/O.

Selection of Salary > Rs. 40000 is also applied on the fly

The selection on salary and the projection on Ename and designation are then applied on the fly at no additional cost.

Hence the total cost => 8 I/O + 24 I/O => 32 I/O.

****