Department of Mathematics
IIT Kharagpur

End Spring Semester Examination 2017-2018
Subject: File Organization & Database Systems(MA40004/MA61018/MA60050)
4th Yr. B. Tech., 4th Yr. M.Sc., 1st Yr. M. Tech (CSDP)
No. of Students (132)

Time: 03 Hours Marks: 50

Answer all FIVE Questions
(This question paper consists of THREE Pages)

Q. 1. a) Consider the following schedule over transactions T_1, T_2, T_3:

T_1	T_2	T_3
	read (Z)	
	read (Y)	
	write (Y)	
		read (Y)
		read (Z)
read (X)		
write (X)		
		write (Y)
		write (Z)
	read (X)	
read (Y)		
write (Y)		
	write (X)	

(i) Draw the precedence graph and check whether the schedule is serializable or not.

(ii) Design a concurrent schedule of T_1, T_2 and T_3 that is serializable. Specify also the equivalent serial schedule.

(iii) Assume the timestamp order protocol $TS(T_1) < TS(T_2) < TS(T_3)$. Is the schedule in (a) above possible under timestamp ordering protocol? Justify your answer.

b) Define two phase locking protocol. Prove that two phase locking protocol ensures serializability of a schedule. **(2M+3M+3M+2M)**

Q. 2. a) Consider the following relation :

Course	StuID	Grade
C1	2100	18
C2	2157	18
C2	2230	30
C1	2177	24
C3	2340	30
C2	2200	23
C1	2157	28
C4	2300	30
C1	2263	25
C4	2299	28

Show the following index structures and file organisations:

(i) An indexed-sequential file organisation with a primary sparse index on **StuID**. For a search-key k, an index entry is created if **k mod 100 = 0**.

Contd...

(ii) On the top of the indexed- sequential structure in (i), a secondary B^+ tree index on **Grade**. Assume that the order of B^+ tree is 1. The tuples are read sequentially as stored in the indexed- sequential file in (a) above.

(iii) A hash file organisation using extendable hashing on **Grade** and the hash function **h(v) = v mod 8**. Each bucket holds at most 2 tuples. Show the structure after inserting first five tuples and after inserting all tuples.

b) For a heap file organisation, give estimates of (i) time to fetch a record, (ii) time to insert a record in terms of **r**, **s**, **btt** and **ebt**. For what type of operations heap file organisation is most suitable? Justify your answer. **(2M+3M+3M+2M)**

Q. 3. a) Consider the following grocery store example :

Transaction ID	Items purchased
T1	H1, B1, K1
T2	H1, B1
T3	H1, C1, C2
T4	C2, C1
T5	C2, K1
T6	H1, C1, C2

Use **Apriori** Algorithm to find the candidate and frequent item sets for each database scan. Enumerate all the final frequent item sets. Also derive the association rules that are generated. Assume that minimum support threshold s = 30% and minimum confidence threshold c = 60%.

b) Distributed DBMS are based on various architectures. Using a diagram describe the reference architecture of a distributed DBMS.

c) Distributed databases can be fragmented in many ways. Define the term fragmentation and explain using real world examples any three types of fragmentation that can be carried out.

(3M+3M+2M+2M)

Q. 4. a) Given the following three linked tables in which primary key columns are underlined :

Customer (C-ID, name, country)

Products (prodID, price)

Orders (orderID, C-ID, prodID, o-date)

and the following query :

```
Select Customer.name
from Customer, Orders, Products
where Customer.C-ID = Orders.C-ID
and Orders.prodID = Products.prodID
and Orders.o-date = '10/04/2018'
and Products.price > 100;
```

Contd...

Suppose this query is run by executing the following sequence of steps:

1. R1 = Join of Customer and Orders
2. R2 = Join of Products and R1
3. R3 = Selection (date = '10/04/2018') from R2
4. R4 = Selection (price > 100) from R3
5. R5 = Projection (name) from R4

i) What kind of problem(s) will be faced if the query is executed based on the sequence above.

ii) Suggest a new sequence that will make the query more efficient. You should introduce extra steps and not simply re-arrange existing steps.

b) With reference to the database

Suppliers (S_no, S_name, S_city), Parts (p_no, p_name, city) and Shipments (S_no, p_no, quantity) :

express the following queries in **SQL and QUEL**

i) Find the name of suppliers who supplied maximum quantity of part 'P2'.

ii) Find all the part names (p_name) supplied by supplier 'S2'.

(3M+3M+2M+2M)

Q. 5. a) Discuss ID3 decision tree algorithm. Use this algorithm to construct the decision tree for the data set mentioned below :

Outlook	Temperature	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Note that we wish to predict the value of **Play** using Outlook, Temperature, Humidity and Wind.

b) In the process of recovery of a database from a system failure, discuss the use of log table. Define (i) transaction-consistent check point, (ii) transaction-oriented check point and (iii) write-ahead log strategy.

c) What is deadlock? How deadlock can be detected? Explain with one example. **(5M+3M+2M)**

---------- X ----------