MASTER OF COMPUTER SC. & ENGG. EXAM. - 2018
(1st Semester)
ADVANCED DATABASE SYSTEM CONCEPTS

Time: Three Hours **Full Marks : 100**

Answer any *five* questions.

1.(a) What is a *recoverable* schedule? Why is recoverability of schedules desirable? Are there any circumstances under which it would be desirable to allow non-recoverable schedules. 5

(b) What is a *cascadeless* schedule? Show that every *cascadeless* schedule is also recoverable. 5

(c) Discuss about the possible transaction states along with state transition diagram. 5

(d) What are the reasons for which buffer managers use a steal and no-force approach? 5

2. (a) How can starvation be avoided in two-phase locking protocol? What is rigorous two-phase locking? 5

(b) What is Phantom phenomena? Discuss how the Index locking protocol takes care of the Phantom phenomena. 6

(c) Discuss whether *Dirty Read*, *Unrepeatable Read* and *Phantom phenomena* are possible if the isolation level of a transaction is READ COMMITTED. 6

(d) In timestamp ordering, W-timestamp(Q) denotes the largest timestamp of any transaction that executed write(Q) successfully. Suppose that, instead, we defined it to be the timestamp of the most recent transaction to execute write(Q) successfully. Would this change in wording make any difference? Explain your answer. 3

3. (a) The degree of durability in a remote backup system can be classified as One-safe, Two-very-safe and Two-safe. What are these classes? 3

(b) Consider the following log records in the log file.

LSN	prevLSN	transID	type	pageID	length	offset	before	after
10	NULL	T1000	update	P500	3	21	ABC	DEF
20	NULL	T2000	update	P600	3	21	HIJ	KLM
30	20	T2000	update	P500	3	21	GDE	QRS
40	10	T1000	update	P505	3	21	TUV	WXY

Show the contents of the Transaction Table and the Dirty Page Table at the time when all the log records have been scanned after recovery. 6

(c) What is fuzzy checkpointing and why is it done? 6

(d) Why is it necessary to carry both redo and undo in the ARIES recovery algorithm? What is the order in which these two operations are carried out and why? 5

4. (a) What is the blocking problem in the two phase commit protocol? How the blocking problem can be solved by including a list of all subordinates in the *prepare* message? 5

(b) Suppose that the coordinator includes a list of all subordinates in the *prepare* message. The coordinator fails after sending out either an *abort* or *commit* message. How the active sites can terminate the transaction without waiting for the coordinator to recover? Assume that some but not all of the *abort / commit* messages from the coordinator are lost. 5

(c) Consider that a subtransaction in a global transaction does no updates. How the site in which the subtransaction is running will respond when the two phase commit protocol is executed by the transaction coordinator at the site where the global transaction was initiated? 5

(d) Discuss the biased protocol in a distributed database management system. How unique global timestamps can be generated in a distributed database systems? 5

5. (a) Describe the TF-IDF approach of measuring the relevance of a document to a query in an Information Retrieval system. 5

(b) How do you define the *precision* and *recall* measures in an Information retrieval system? How do you calculate the above evaluation measures on the basis of the following statistics – true positive, truce negative, false positive and false negative? 5

(c) What is an inverted index? How inverted index can be used to allow proximity based ranking? How inverted index can be used for handling 'and', 'or' and 'not' queries in an information retrieval system? 5

(d) What is stemming and why is stemming used in Information Retrieval systems? What is Hub and Authority based Ranking? 5

6. (a) How data can be collected in a data warehouse? Give an example of Star Data Warehouse schema. 6

(b) What are the measures of purity of a set S of training instances? What is information gain ratio? 6

(c) Describe the *support* and *confidence* measures of association rules using appropriate examples. 5

(d) How does the naïve Bayesian classifiers find the probability of an instance d being in a class c_j? 3

7. (a) Consider a Library Information System with a relation *Books* whose attributes are as shown below:

title, a list (array) of authors, Publisher with subfields *name* and *branch* and a set of keywords.

Define the above scheme in SQL with appropriate types for each attribute. 6

(b) Write the SQL statement for insertion of a book tuple with the following values: title='Compilers', authors='Smith', 'Jones', Publisher='Mc-Graw Hill, New York', Keywords = 'parsing', 'analysis'. 2

(c) Write SQL statements to accomplish the following tasks: 4

(i) to find all books that have the word "database" as a keyword.

(ii) To get a relation containing pairs of the form "title, author_name" for each book and each author of the book

(d) What is nesting? Discuss with a suitable example. 6

(e) Define a type Department with a field name and a field head which is a reference of the type Person, with table people as scope. 2

8. (a) What are the three broad levels at which a database system can be tuned to improve performance? Give two examples of how tuning can be done for each of the levels. 9

(b) What is the motivation for splitting a long transaction into a series of small ones? What problems could arise as a result, and how can these problems be averted? 4

(c) Suppose that a database application does not appear to have a single bottle-neck, that is CPU and disk utilization are both high, and all database queues are roughly balanced. Does that mean that the application cannot be tuned further? Explain your answer. 3

(d) Suppose a system runs three types of transactions. Transactions of type A run at the rate of 50 per second, transactions of type B run at 100 per second and transactions of type C run at 200 per second. Suppose the mix of transactions has 25 percent of type A, 25 percent of type B and 50 percent of type C.

What is the average transaction throughput of the system, assuming there is no interference between the transactions?

What factors may result in interference between the transactions of different types, leading the calculated throughput being incorrect? 4

---------- XX ----------