UNIVERSITY OF PERADENIYA
Faculty of Engineering

END OF SEMESTER EXAMINATION, JULY 2020

CO 527 ADVANCED DATABASE SYSTEMS
(Three Hours)

Registration Number:	**E/...../.....**
1. Answer **all** questions on this booklet itself. 2. All questions carry equal marks.	

For Examiner's Use Only							
	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Total**
Marks Awarded							

1. Query Optimization and Database Tuning

a. What is a query execution plan? [10 Marks]

b. Discuss the main heuristics that are applied during query optimization. [10 Marks]

c. Consider the following relations and the SQL query based on the relations.

Relations:

EMPLOYEE (Ssn, Fname, Lname, Address, Bdate)

DEPARTMENT (Dnumber, Dname, Mgr_ssn)

PROJECT (Pnumber, Pname, Plocation, Dnum)

SQL query:

```
SELECT Pnumber, Dnum, Lname, Address, Bdate
FROM ((PROJECT JOIN DEPARTMENT ON Dnum=Dnumber)
        JOIN EMPLOYEE ON Mgr_ssn=Ssn)
WHERE Plocation='Stafford';
```

i. Draw an initial query tree to represent the query given above. [10 Marks]

ii. Show how the initial query tree is optimized based on the heuristics mentioned in (1.b) and find a final query tree that is efficient to execute. Briefly explain the steps. [30 Marks]

d. List three factors that influence physical database design. Briefly explain how those factors could be used to revise database design. [20 Marks]

e. What are the reasons for tuning indexes when improving the overall throughput of transactions? [20 marks]

2. Transaction Processing

a. What is meant by strict schedule? [20 marks]

b. Is the following schedule with three transactions T_1, T_2 and T_3 operating on three variables X, Y and Z a strict schedule? Explain. [20 Marks]

$R_1(X)$, $R_2(Z)$, $R_1(Z)$,$R_3(X)$, $R_3(Y)$, $W_1(X)$, c_1 (Commit), $W_3(Y)$, c_3 (Commit), $R_2(Y)$, $W_2(Z)$, $W_2(Y)$, c_2(Commit);

c. What is a serializable schedule? [20 Marks]

d. Show that the following schedule with three transactions T_1, T_2 and T_3 operating of three variables X, Y and Z is not serializable.

$R_2(Z)$, $R_2(Y)$, $W_2(Y)$, $R_3(Y)$, $R_3(Z)$, $R_1(X)$, $W_1(X)$, $W_3(Y)$, $W_3(Z)$, $R_2(X)$, $R_1(Y)$, $W_1(Y)$, $W_2(X)$ [20 Marks]

e. Can the schedule in 2 (d) be made serializable? If so, how? [20 Marks]

3. Concurrency Control and Database Recovery Techniques

a. How does two phase locking protocol guarantee serializability? [20 Marks]

b. Show that the following schedule does not satisfy the two-phase locking protocol. [20 Marks]

T_1: Read_Lock (Y), T_1: Read_Item (Y), T_1: Unlock (Y), T_2: Read_Lock (X), T_2: Read_Item (X), T_2: Unlock (X), T_2: Write_Lock (Y), T_2: Read_Item (Y), T_2: Write_Item (Y), T_2: Unlock (Y)

c. What is snapshot isolation? What are the advantages and disadvantages of concurrency control methods that are based on snapshot isolation? [20 Marks]

d. Discuss the different types of transaction failures. What is meant by catastrophic failure? [20 Marks]

e. Describe the three phases of ARIES recovery method. [20 Marks]

4. Distributed Database Systems

a. Why is the performance of distributed database system better compared to a centralized system? [10 Marks]

b. Giving an example, explain horizontal fragmentation in distributed database systems. [20 Marks]

c. What are the challenges in handling transactions in distributed database systems? [10 Marks]

d. Explain how three-phase commit protocol handles the issues of two-phase commit protocol. [20 Marks]

e. Consider the data distribution of the BOOK database, where data is fragmented at two sites as follows.

Site 1: BOOK (BookID, Title, Price, AuthorID)

10,000 records

Size of the fields: BookID: 10 bytes, Title: 35 byes, Price: 5 bytes, AuthorID: 5 bytes

Site 2: AUTHOR (AuthorID, Name, Address, Telephone)

2,000 records

Size of the fields: AuthorID: 5 bytes, Name: 15 bytes, Address: 20 bytes, Telephone: 10 bytes

Assume that the following query Q is submitted at site 3, which is distinct from site 1 and 2.

$$\text{Q: } \pi_{Title,Name}(BOOK * AUTHOR)$$

iii. Identify the best strategy for executing the query by considering optimum criterion as the amount of data transferred between sites. Justify your choice. [30 Marks]

iv. Suppose that query Q is submitted at site 2. Identify the best strategy for executing the query. [10 Marks]

5. Big Data

a. Define “big data”. [10 Marks]

b. Briefly explain three example applications which use big data. [30 Marks]

c. Providing examples, describe the characteristic *Variety* in big data. [10 Marks]

d. What is Apache Hadoop? [20 Marks]

e. List the components of the Hadoop ecosystem with a brief description. [30 Marks]

6. NoSQL databases

a. For which types of applications were NoSQL systems developed? [20 Marks]

b. What are the main categories of NoSQL systems? Provide an example for each category. [20 Marks]

c. What are the main characteristics of NoSQL systems in the areas related to distributed systems and distributed databases? [20 Marks]

d. What is the CAP theorem? Which of the three properties are most important in NOSQL systems? [20 Marks]

e. Briefly explain MongoDB data model. [20 Marks]

End of the Paper