Xi'an Jiaotong-Liverpool University
西交利物浦大学

XJTLU Entrepreneur College (Taicang) Cover Sheet

Module code and Title	Database Development and Design (DTS207TC)	
School Title	School of AI and Advanced Computing	
Assignment Title	002: Assessment Task 2 (CW)	
Submission Deadline	23:59, 12th Dec (Friday)	
Final Word Count	NA	
If you agree to let the university use your work anonymously for teaching and learning purposes, please type **"yes"** here.		**Yes**

I certify that I have read and understood the University's Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on Learning Mall Online). With reference to this policy I certify that:

- My work does not contain any instances of plagiarism and/or collusion.
- My work does not contain any fabricated data.
- My work does not contain any text or code from the Internet, Generative AI, or published textbooks

By uploading my assignment onto Learning Mall Online, I formally declare that all of the above information is true to the best of my knowledge and belief.

Scoring – For Tutor Use	
Student ID	

Stage of Marking	**Marker Code**	**Learning Outcomes Achieved (F/P/M/D)** **(please modify as appropriate)**			**Final Score**
		A	**E**		
1st Marker – red pen					
Moderation – green pen	**IM Initials**	The original mark has been accepted by the moderator (please circle as appropriate):			Y / N
		Data entry and score calculation have been checked by another tutor (please circle):			Y
2nd Marker if needed – green pen					

For Academic Office Use			**Possible Academic Infringement (please tick as appropriate)**	
Date Received	**Days late**	**Late Penalty**	□ **Category A**	Total Academic Infringement Penalty (A,B, C, D, E, Please modify where necessary) ________________
			□ **Category B** □ **Category C**	

			□ **Category D**	
			□ **Category E**	

Xi'an Jiaotong-Liverpool University
西交利物浦大学

Database Development and Design (DTS207TC)

Assessment 002: Individual Coursework

Weight: 40%

Maximum Marks: 100

Overview & Outcomes

This course work will be assessed for the following learning outcomes:

A. Identify and apply the principles underpinning transaction management within DBMS.
E. State the main concepts in data warehousing and data mining.

Submission

You must submit the following files to LMO:

1)A report named as Your_Student_ID.pdf.

2)A directory containing all your source code, named as Your_Student_ID_code.

NOTE: The report shall be in A4 size, size 11 font, and shall not exceed **9** pages in length. You can include only key code snippets in your reports. The complete source code can be placed in the attachment.

Question 6: Storage Management (40 marks)

In a database storage system, the cache hit rate has a significant impact on its performance. Different cache strategies will result in different cache hit ratios. Now, we have recorded 2 datasets (please download from LMO), containing CPU access requests to memory for a period of time. They both have 10,000 items from addresses 0 to 63. We will simulate the process of the CPU reading and caching data from the memory through a program in the table below (also can be download from LMO) . Please run the program to compare the hit rates of different strategies:

Python

```
import random
from collections import deque

class RandomPolicy:
    def __init__(self, size):
        self.size = size
        self.cache = []
        self.name = 'rr'

        random.seed(207)
    def access(self, current):
        if current in self.cache: # hit!
            return True

        self.cache.append(current)

        if len(self.cache) > self.size: # exceed
            self.cache.remove(random.choice(self.cache))

        return False


class FifoPolicy:
    def __init__(self, size):
        self.size = size
        self.cache = deque()
        self.name = 'fifo'

    def access(self, current):
        if current in self.cache: # hit!
            return True

        if len(self.cache) == self.size: # full
            self.cache.popleft()

        self.cache.append(current)

        return False


def run_test(trace, pol):
    hit = []
    for i in range(len(trace)):
        # update cache
        hit += [pol.access(trace[i])]
    return sum(hit) / len(hit)


if __name__ == '__main__':
    # parameters
    caps = [1, 2, 3, 4, 5]

    # load trace from file
    traces = []
    for name in ['trace1.txt', 'trace2.txt']:
        with open(name) as f:
            traces += [list(eval(f.read()))]

    # test all strategies
    strategies = [
        FifoPolicy,
        RandomPolicy,
    ]

    # run strategy over trace
    for i in range(len(traces)):
        for cap in caps:
            for Strategy in strategies:
                pol = Strategy(size=cap)
                print(f'data={i +
1},\tcap={cap},\tname={pol.name},\thitrate={run_test(traces[i], pol)}')
            print()
```

You need to analyze the characteristics of this data and analyze why the hit rates of the two strategies are different on the two data sets (20 mark). Design and implement a strategy which can achieve better results than the **RandomPolicy** strategy on the **trace2** data set. Record the hit rates you observed in the table below (with snapshot) (20 marks).

Cache Size	RR	Your Policy
1		
2		
3		
4		
5		

Question 7: Indexing (30 marks)

Consider a hard disk with a sector size of B = 512 bytes. A CUSTOMER file contains approximately r = 40,000 records. Each record includes the following fields: Name (30 bytes), Ssn (9 bytes), Email (30 bytes), Address (50 bytes), Phone (15 bytes), and Birth_date (8 bytes). The Ssn field is the primary key. The file system uses 4KB blocks for allocation.

(a) Calculate the number of blocks required for an unspanned organization. Then, discuss how the discrepancy between sector size and block size might affect sequential access performance, and whether you would recommend using a different block size for this scenario. (6 marks)

(b) The records are physically ordered on Ssn. Calculate the maximum number of block accesses for a binary search. During system testing, developers notice that batch queries processing large ranges of Ssn values perform 30% slower than expected when using binary search as the primary lookup method. Provide two possible explanations for your scenario. (6 marks)

(c) A sparse index is built on Ssn. Calculate the number of block accesses to retrieve a record using this index in ideal scenario. During usage, it is found that the performance of the index search continues to decline. Identify two potential reasons why the index performance gain is less than theoretical expectations. (6 marks)

(d) A multi-level primary index is constructed. During the design review, two proposals are made: Proposal A: Use the standard multi-level index structure; Proposal B: Based on Proposal A, select 10 index blocks and cache them persistently in memory. Calculate the number of index levels needed for the Proposal A. Then, compare the two proposals in terms of implementation complexity and computation time under a workload with 10% of the Ssn accounting for 90% of the queries. (6 marks)

(e) A B+ tree index is built with order p = 50. Calculate the maximum number of records a height 4 tree can index. During maintenance, it's observed that, during frequent insertion and deletion operations, the tree height changes frequently between 3 and 4 even with relatively stable data size. Explain what might be causing this fluctuation and suggest one strategy to stabilize the tree height with optimal performance. (6 marks)

Question 8: Transaction (30 marks)

Consider a database with a relation *Account (AccountID, Balance)* and initial state:

AccountID	Balance
1	110
2	10

(a) The following transactions represent a fund transfer (T1) and a real-time balance report (T2) that run concurrently. (10 marks)

T1 (Transfer)	T2 (Report)
1. begin transaction 2. update Account set Balance = Balance - 100 where AccountID = 1; 3. update Account set Balance = Balance + 100 where AccountID = 2; 4. commit;	1. begin transaction 2. select sum(Balance) from Account; 3. commit;

The application requirement states that the report must never reflect a financially inconsistent state. However, under certain database configurations, T2 might output a total of 20 (instead of the correct 120).

(i) Explain under which isolation level(s) this inconsistent total of 20 could occur, and describe the exact sequence of operations in a concurrent schedule that leads to this result.

(ii) The development team proposes using the SERIALIZABLE isolation level to fix this issue. Critically evaluate this proposal by discussing one key advantage and two potential drawbacks (considering both performance and system complexity) for this specific application scenario.

(b) Now consider these transactions: a process adding a new account (T3), and an audit process calculating the total balance (T4). (10 marks)

T3(New Account)	T4 (Audit)
1. begin transaction 2. insert into Account values (3, 150); 3. commit;	1. begin transaction 2. select sum(Balance) from Account; 3. select sum(Balance)

	from Account; 4. commit;

Suppose the application requirement for the audit is that it must have a consistent view of the database throughout its execution.

(i) Is it possible for the two SUM queries in T4 to return different values? Analyze this possibility under at least three different isolation levels, providing a brief concurrent schedule for each case where the results differ.

(ii) During a system design review, an engineer suggests: "We can just use the REPEATABLE READ isolation level to solve all our concurrency problems in this audit process." Write a brief response evaluating this suggestion. Your response should consider whether this is sufficient, necessary, and practical for meeting the audit requirement.

(c) Consider the following observation from production logs: (10 marks)

- T5 (Data Maintenance): Inserts 100 new account records in a single transaction.
- T6 (Analytics Query): Runs *SELECT COUNT(*) FROM Account*; twice within its transaction and gets two different results.

The team initially diagnosed this as a "phantom read."

(i) Under which isolation level(s) is this phenomenon possible?

(ii) A DBA comments: "While this looks like a phantom read, the actual impact and the appropriate fix might be different if those 100 new accounts were all inserted with a zero balance." Briefly explain the DBA's reasoning. Why might the business impact and the technical solution be different if the new accounts have a zero balance, even though the phenomenon looks the same?

Marking Criteria

The tasks in this assessment can be divided into 3 categories:

- ✓ Charts Presentation & Analysis;
- ✓ Essay;
- ✓ Programs.

Criteria(%)	Exemplary (100)	Good (75)	Satisfactory (50)	Limited (25)	Very Limited (0)
Design	Provides a detailed, accurate description of	The analysis provided demonstrates that the	Provides adequate description of the methods.	There are obvious deviations in the understanding of the main	Limited or no description of methods. Limited

	the methods. Provide comprehensive comparison between the methods, including pros and cons, performance analysis.	student's understanding of the various methods is correct and that they have the ability to solve problems independently. Although there are certain flaws, or incomplete.	Comparison is provided with some level of details, however, with some obvious mistakes.	methods, and it fails to reflect the ability to independently design algorithms. The description of the problem is vague, or the thought is incomplete.	comparison provided.
Programs	Demonstrated correctly implemented code that produces correct output. Excellent coding quality follows best practices.	The program runs correctly and gives the expected results. However, special cases are not fully considered, or the program performs redundant calculations.	Program basically works correctly for major functionality, however, with some conceptional problems.	The program implements some minor functionality, or incorrectly implements major functionality. There is a certain degree of misunderstanding about the requirements of the questions.	Program works incorrectly with limited attempt or irrelevant to the task.
Charts Presentation & Analysis	Excellent quality of report with clear structure, clear logic, concise writing, pleasing visual aids.	Most of the results in the chart are correct, but there is a certain degree of sloppy or wordy in the overview and analysis.	Moderate quality of report with basic structure, where writing and visual aids can be improved.	Only some of the results in the chart are correct, or some of them are not filled in. The analysis of the results was obviously biased.	Limited or no attempt of report.

The mark allocations for the above tasks are:

Task	Design	Programs	Charts Presentation & Analysis
6	25	10	5
7	30		

8	30		