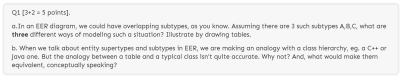
My grades for **Midterm exam**

Q1

0 / 5



Q1 [3+2 = 5 points].

a.In an EER diagram, we could have overlapping subtypes, as you know. Assuming there are 3 such subtypes A,B,C, what are **three** different ways of modeling such a situation? Illustrate by drawing tables.

b. When we talk about entity supertypes and subtypes in EER, we are making an analogy with a class hierarchy, eg. a C++ or Java one. But the analogy between a table and a typical class isn't quite accurate. Why not? And, what would make them equivalent, conceptually speaking?

This question wasn't answered

Q2

5 / 5

Q2 [1+4 = 5 points].

a. What is the benefit of normalization, what is its drawback?

b. In a class of students, each student has an ID and a name. Each student is assigned (given) a book by a popular author to read; many students could be assigned the same book (eg. many might be assigned to read 'The Adventures of Tom Sawyer', by Mark Twain). The class teacher uses a spreadsheet to keep track of the # of hours a student puts in, towards reading her/his book.

Show, using a table, how the teacher would store data incorrectly. Show how you would help fix the table. To save time when you answer, you can use 'simple' values like A,B,C... for your data (they don't need to be 'real').

Q.2)

a.) Normalization has the advantage of reducing data redundancy and enhancing data integrity, which makes it simpler to maintain and update the database. We can lessen data anomalies, enhance data consistency, and save storage space by splitting huge tables into smaller ones and deleting redundant data. Normalization has a disadvantage in that it necessitates joining many tables in order to get data, which can possibly slow down query execution.

b.)

Student ID	Student Name	Book Title	Author	Hours Read
A	Alice	Tom Sawyer	Mark Twain	10
B	Bob	Tom Sawyer	Mark Twain	12
C	Cindy	Tom Sawyer	Mark Twain	8
D	Dave	Treasure Island	Robert Louis Stevenson	9

The Book Title and Author columns in this table contain duplicate information, which may cause data anomalies if the same book is spelled differently or if the author's name is misspelled. Also, the teacher will need to alter numerous rows if they want to modify the book's title or author, which can be time-consuming and error-prone.

A better table shall be:

Student ID	Student Name	Book ID	Hours Read
A	Alice	1	10
B	Bob	1	12
C	Cindy	1	8
D	Dave	2	9

Book ID	Book Title	Author
1	Tom Sawyer	Mark Twain
2	Treasure Island	Robert Louis Stevenson

According to this approach, each book is given a distinct ID that serves as a foreign key in the student table. The book information is kept in a separate database. This gets rid of unnecessary information and makes it simple to update the material in the book without impacting the student data.

Q3

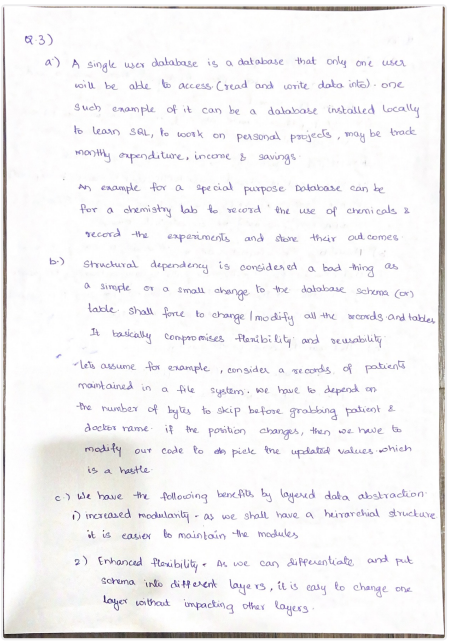
5 / 5

Q3 [1+2+2 = 5 points].

a. What is an example of data that is suitable for a single-user DB? What is another example, for a special-purpose DB?

b. Why is structural dependence a bad thing, when it comes to storing data? Illustrate structural dependence using a small example of your own [with some sample data].

c. On the flip side (of structural dependence), we have layered data abstraction - what benefit does layering provide? Explain in two or three sentences (NOT more!).



Q 3)

a) A single user database is a database that only one user will be able to access (read and write data into) - one such example of it can be a database installed locally to learn SQL, to work on personal projects, may be track monthly expenditure, income & savings.

An example for a special purpose database can be for a chemistry lab to record the use of chemicals & record the experiments and store their outcomes.

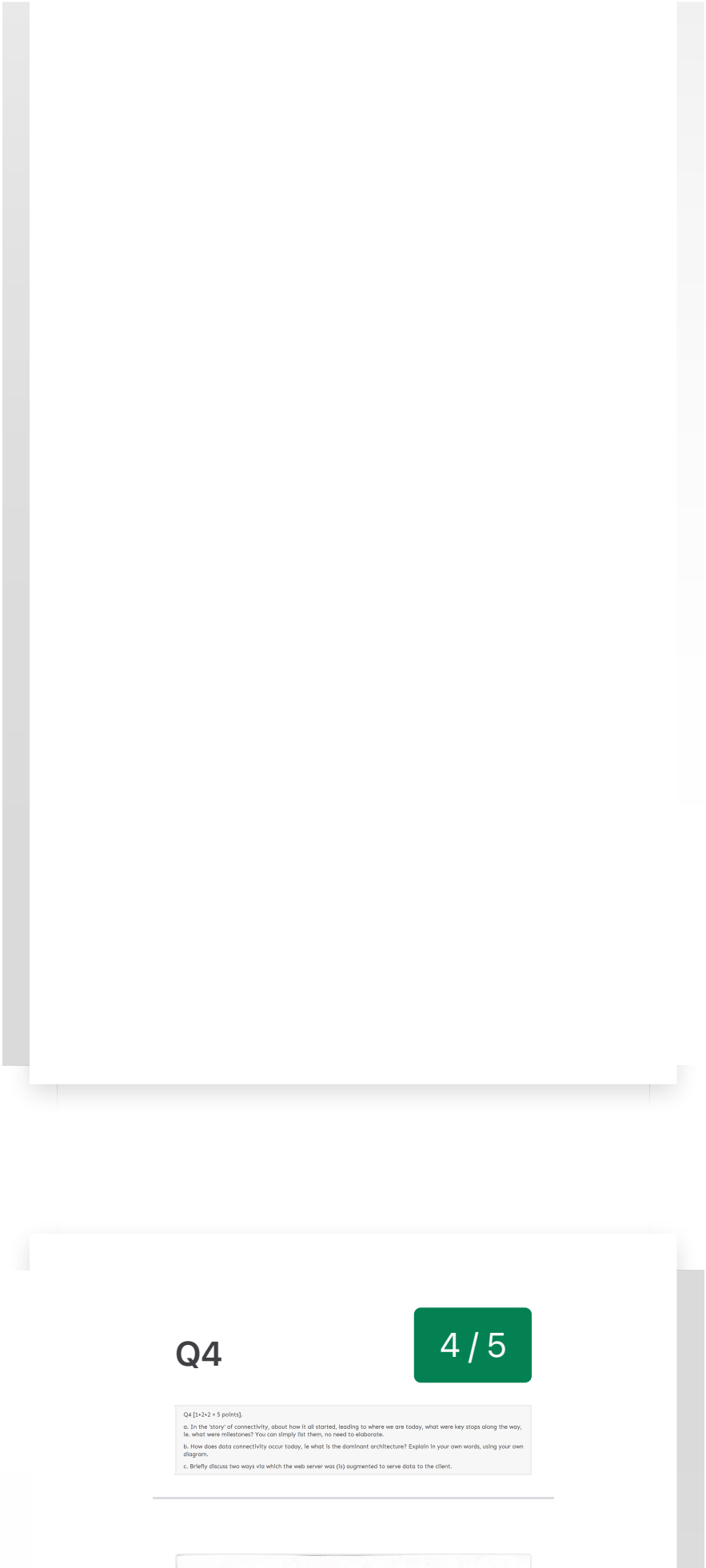
b) Structural dependency is considered a bad thing as a simple or a small change to the database schema (ex) table shall force to change / modify all the records and tables. It basically compromises flexibility and reusability.

Let's assume for example, consider a records of patients maintained in a file system. We have to depend on the number of bytes to skip before grabbing patient & doctor name. if the position changes, then we have to modify our code to also pick the updated values which is a hastle.

c) We have the following benefits by layered data abstraction.

1) increased modularity - as we shall have a hierarchial structure it is easier to maintain the modules.

2) Enhanced flexibility - As we can differentiate and put schema into different layers, it is easy to change one layer without impacting other layers.



Q4

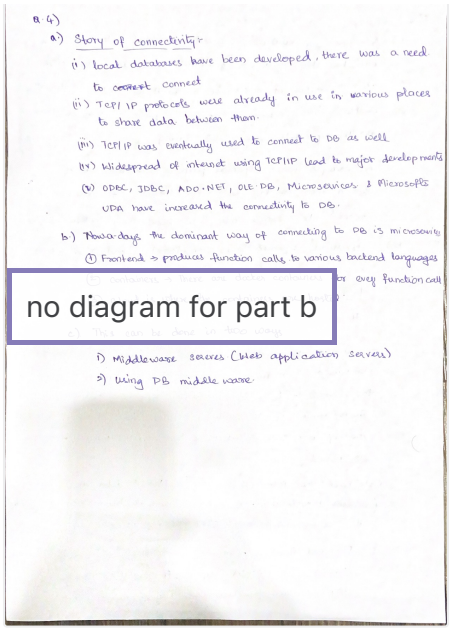
4 / 5

Q4 [1+2+2 = 5 points].

a. In the 'story' of connectivity, about how it all started, leading to where we are today, what were key stops along the way, ie. what were milestones? You can simply list them, no need to elaborate.

b. How does data connectivity occur today, ie what is the dominant architecture? Explain in your own words, using your own diagram.

c. Briefly discuss two ways via which the web server was (is) augmented to serve data to the client.



q 4)

a) Story of connectivity:-

i) local databases have been developed, there was a need to connect connect

ii) TCP/IP protocols were already in use in various places to share data between them.

iii) TCP/IP was eventually used to connect to DB as well

iv) Widespread of internet using TCP/IP lead to major developments

v) ODBC, JDBC, ADO.NET, OLE DB, Microservices & Microsoft UDA have increased the connectivity to DB.

b) Nowadays the dominant way of connecting to DB is microservices

① Frontend → produces function calls to various backend languages

every function call

no diagram for part b

c) This can be done in two ways

i) Middleware servers (Web application servers)

ii) Using DB middle ware

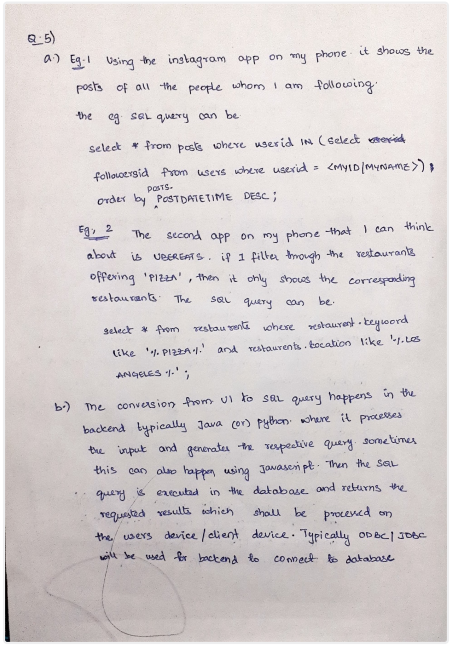
Q5

5 / 5

Q5 [4+1 = 5 points].

a. Pick any two apps/sites on your phone/table/laptop using which you access data, describe how your UI actions (eg. searching, or doing data filtering) might result in SQL, using one example for each app/site (so two examples total).

b. Assuming (like in 'a' above) that your app-driven-querying does turn into SQL, where would such conversion (ie. transformation from UI-based query to SQL) occur?



Q.5)

a.) Eg.1 Using the instagram app on my phone it shows the posts of all the people whom I am following.

the eg. SQL query can be

```
select * from posts where userid IN (select ~~userid~~
followersid from users where userid = <MYID/MYNAME>) ;
order by POSTDATETIME DESC ;
```

(posts.)

Eg. 2 The second app on my phone that I can think about is UBEREATS. if I filter trough the restaurants offering 'PIZZA', then it only shows the corresponding restaurants. The SQL query can be.

```
select * from restaurants where restaurant.keyword
like '%PIZZA%' and restaurants.location like '%LOS
ANGELES%' ;
```

b.) The conversion from UI to SQL query happens in the backend typically Java (or) Python. where it processes the input and generates the respective query sometimes this can also happen using javascript. Then the SQL query is executed in the database and returns the requested results which shall be processed on the users device/client device. Typically ODBC/JDBC will be used for backend to connect to database

Q6

5 / 5

Q6 [2+2+1 = 5 points].

a. In 2PL for data access during transactions, what is the most important phase? Explain.

b. In 2PL, when we release locks, if we release locks prematurely, what issue might that cause? How would we fix it?

c. What issue might arise, when we do TM without locks? How would we fix it?

Q.6)

a.) The lock acquisition phase in 2PL is very crucial for accessing the data during transactions. The transaction requests and obtains all the necessary locks on the data items it needs to access and modify during this phase. This phase shall make sure that until the first transaction is finished, additional transactions cannot access or modify the same data items. If this is not done correctly, then the transaction may produce inconsistent or inaccurate data and break the isolation principle of the ACID properties for transactions.

b.) Prematurely releasing of the locks in 2PL can result in lost-updates scenario. This happens when the data item is updated by one transaction, but also the same data item is updated by another transaction before the first transaction has done a COMMIT on its modifications. The original transaction's modifications are thereby undone, and the data is now inconsistent. To avoid this situation, we can utilize a strict two-phase locking, which shall keep all locks in place until the transaction has either done a COMMIT or a ROLLBACK.

c.) Withouttheuseoflocks,transactionmanagementcanleadtodirtyreadswhichisa problem. When a transaction reads a piece of data that has already been updated by another transaction but hasn't been committed, this happens. It may produce inconsistent or inaccurate outcomes when the initial transaction bases its operations on the updated (but uncommitted) data. We can use optimistic concurrency control to resolve these types of problem, which entails allowing transactions to read and write data without restriction while checking for conflicts at commit time. All transactions commit only if they do not conflict with one another, so if a conflict is found, one of the transactions is rolled back & resumed.

Q7

5 / 5

Q7 [1+4 = 5 points].

a. Codd's relational operators for data processing, lead to 'closure'. Why is this advantageous? Illusrate.

b. For these four data types, list an operation that does preserve closure, and one that does not: vector (eg with components x,y,z), matrix, complex number, color (with RGB components).

Q.7.)

a.) The closure of Codd's relational operators for data processing results in a relation that can be used as input to another operator after the operators have been applied to a set of relations. This has benefits since it makes data processing effective and modular. If we have two relations A and B, for instance, we may use the selection operator on each relation to produce the subsets A' and B' before joining the two subsets to create a new relation C. The data can then be processed further by using additional operators on C, such as projection, grouping, or sorting. We can combine the operators in a variety of ways to get the desired outcomes since they maintain closure.

b.) Vector: Scalar multiplication, in which the vector is multiplied by a scalar quantity, is an operation that maintains closure. Taking the cross product of two vectors, which yields a vector that might not be orthogonal to the source vectors, is an operation that does not preserve closure.

Matrix: Matrix addition, which involves adding two matrices element-by-element, is an operation that protects closure. Matrix inversion is an operation that does not preserve closure and may produce a non-invertible matrix.

Complex Number: Addition, which involves combining two complex numbers component-by-component, is an operation that maintains closure. Taking a complex integer's square root is an operation that does not maintain closure, and the outcome can be a complex number with a non-real component.

Color: Color addition, which involves combining two colors component-by-component, is an operation that keeps closure. Color multiplication is an operation that does not preserve closure and may produce a color that is not in the RGB color space.

Q8

4 / 5

Q8 [2+2+1 = 5 points].

a. During 2PC in distributed transactions, the transaction coordinator might fail. How would we fix that?

b. During 2PC, a non-coordinator site might fail partway (between phase 1 and phase 2) - how would we fix the problem (ie prevent bad transactions)?

c. During 2PC, a a non-coordinator site might fail at the start (before phase 1) - how would we deal with that?

Q 8)

a.) We can utilize a backup coordinator to assume the role and finish the commit or abort process in distributed transactions when the transaction coordinator fails during the second phase of concurrency control (2PC). In case of failure, the backup coordinator, a selected site with a copy of the transaction log, can take over. When the backup coordinator takes over, it analyzes the transaction log to ascertain the status of the transaction and then moves forward with the required commit or abort actions.

b.) We can avoid bad transactions by employing a timeout mechanism if a non-coordinator site dies midway through 2PC (between phase 1 and phase 2). The coordinator site watches for responses from all non-coordinator sites within a predetermined window of time. The coordinator moves through with the required commit or abort actions for the remaining sites if a non-coordinator site doesn't answer within the timeout period, assuming the site has failed. In this approach, even if one or more non-coordinator sites fail, the transaction can still be properly completed.

c.) A recovery mechanism can be used to handle the scenario if a non-coordinator site fails at the beginning of 2PC (prior to phase 1). The recovery mechanism is in charge of undoing any modifications made by failed transactions and retrieving any lost data in order to return the failed site to a consistent state. The site can take part in the 2PC protocol as usual after it has been restored. The site is withdrawn from the protocol if it cannot be recovered, and the surviving sites carry out the necessary commit or abort operations.

Send a message to all the non-coordinator sites before transaction starts. If any of the sites is not responsive after a certain amount of time, the coordinator can mark that site as failed and schedule the tasks

schedule the tasks so this failed site will not be used in the next time.

Q9

5 / 5

Q9 [1+2+2 = 5 points].

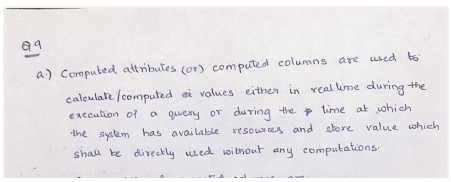
a. What are a couple of uses for 'computed columns'?

b. Given a table with columns of sines and cosines (for 0 to 360 degrees, in increments of 1 degree), eg called COS and SIN, how would you verify the following formula/identity?

$\sin^2\theta+\cos^2\theta=1$

You don't need to write SQL, you can simply describe the steps.

c. Given a table with a pair of columns called X and Y, containing (x,y) values from a scatterplot for example, how would you calculate the (Pearson) correlation coefficient [assuming that the relationship is linear]? Again, just describe the steps [no need for SQL].



Q9

a) Computed attributes (or) computed columns are used to calculate/computed @ values either in real time during the execution of a query or during the ⊕ time at which the system has available resources and store value which shall be directly used without any computations.

1) derive a persons annual income based on his monthly salary
2) derive a persons age based on date of birth
3) combining both first name and last name.

b.) Assuming we are given a table with values of cos & SIN from 0 to 360 with increments of '1'.
We can create a computed column with the formula ((SIN * SIN) + (COS * COS)), and this column shall hold all '1's if the formula $\sin^2\theta + \cos^2\theta$ is true, else it holds a different value.

c.) Calculating Pearson co-relation coefficient from a column of X and Y values we can use the below steps.
1) create the following computed columns.
(i) MEANX = AVG(X)
(ii) MEANY = AVG(Y)
(iii) STDX = SQRT (SUM((X - MEANX) * (X - MEANX))/COUNT(X))
(iv) STDY = SQRT ((SUM (CY - MEANY) * (Y - MEANY))/COUNT(Y))
(v) TEMP = SUM ((X - MEANX) * (Y - MEANY))
(vi) PEARSONCOEFF = TEMP / (STDX * STDY)