Assignment on Database Design for CS50L

Total points 9/10

Starting in 2021, all assignments in CS50L are out of 10 points. A score of 7 points or better (70%) is required to be considered to have "passed" an assignment in this course. Please do not resubmit an assignment if you have already obtained a passing score. You don't receive a final grade at the end of the course, so it will have no bearing on your certificate, and it will only slow down our graders!

Unlike CS50x, assignments in this course are graded on a set schedule, and depending on when you submitted, it may take up to three weeks for your work to be graded. Do be patient! Project scores and assignment status on cs50.me/cs50l (e.g. "Your submission has been received...") will likely change over time and are not final until the scores have been released.

Email *

maz786@outlook.com

Name *

Mazafer Ul-Raqib

edX Username *

Mazafer

What is your GitHub username?

You only need to tell us if you are concerned about checking your progress in the course and/or you want a free CS50 Certificate after you satisfy all of the requirements of the course. If you do not already have a GitHub account, you can sign up for one at https://github.com/join. You can then use this account to log in to cs50.me/cs50l to track your progress in the course (your progress will only show up after you have received at least one score release email from CS50 Bot, so do be patient!). Don't worry about seeing a 'No Submissions' message on submit.cs50.io, if you find that. The course collects submissions using Google Forms, and only the gradebook on cs50.me/cs50l is important! If you do decide to provide us with a GitHub username, BE CERTAIN IT IS CORRECT. If you provide the wrong username, you will not be able to see your scores.

https://github.com/maz786

This course is graded by human graders, and has a ZERO TOLERANCE plagiarism and collaboration policy. If *any* of your answers are copied and pasted from, or obviously based on (a) an online source or (b) another student's work in the course, in *any* of the course's ten assignments, you will be reported to edX and removed from the course immediately. There is no opportunity for appeal. There are no warnings or second chances. *

It is far better, we assure you, to leave an answer blank rather than risk it. This may be an online course, but it is offered by Harvard, and we're going to hold you to that standard.

◉ I understand this policy and agree to its terms; I hereby affirm that I will not plagiarize any answers or collaborate with any other students in this course .

✓ What does it mean for a machine (whether a router or load balancer or database or server more generally) to be a single point of failure? How can it be avoided? And how can avoiding it create new problems? 1/1

A single point of failure (SPOF) is a component in a distributed system that, if it fails, prevents the operation of the entire system. This is a serious problem because a SPOF failure has the ability to bring down the entire system, resulting in serious disruption or loss.

SPOFs can be prevented in a number of ways:

Redundancy: Having many copies of a single component enables the others to take over in the event of a failure.

By dividing the effort among several components, load balancing ensures that no one component is overworked.

Designing a system with modular components that are easily replaceable when one fails is known as modular design.

Avoiding SPOFs, however, can lead to other issues:

Cost: Adding hardware and software might be pricey when implementing redundancy and load balancing.

Redundancy and load balancing increase the system's complexity, which might make it more challenging to administer and maintain.

Performance: As resources are utilised to support the additional hardware and software, redundancy and load balancing can reduce the system's overall performance.

✗ What does it mean to shard a database? And why would you do it? 0/1

Partitioning a database into shards or servers entails distributing the data among them. Usually, this is done to increase the database's scalability and performance.

There are a number of factors that could lead you to wish to shard a database:

Performance improvement: You can distribute the load and improve the database's overall performance by distributing the data across several servers.

Improved scalability: As data volumes grow, managing and storing the data on a single server may become increasingly challenging. When additional servers are required to handle the expanding data size, sharding makes this possible.

High availability: By enabling you to continue processing requests even if one or more servers fail, sharding can also increase the database's availability.

Sharding a database can increase complexity, though, since you now need to manage numerous servers and make sure the data is evenly dispersed and partitioned. It may also cost more since it needs more hardware and software resources.

✓ What does it mean for a database to have a "race condition"? How can we avoid them? 1/1

When two or more processes or threads attempt to access a shared resource concurrently and the results of the processes depend on the order in which they access the resource, this is known as a race condition. This can produce unexpected or inaccurate outcomes.

Here is an illustration of a database race condition:

A reads a value from the database in Process A.
The same value is read from the database by Process B.
Process A modifies the database value.
Process B modifies the database value.
Depending on the sequence in which the processes run in this case, the value in the database can be left in an inconsistent state.

There are a number of techniques to prevent racial situations:

Locking: This entails utilising locks to stop several programmes from simultaneously accessing the same resource.

Transactions: You can join several database operations together into a single transaction and either commit them all at once or roll them back if any of them fail. This guarantees that the database continues to be consistent.

Allowing many processes to utilise the resource concurrently while monitoring for and resolving conflicts as they arise is known as optimistic concurrency control.

Each approach has trade-offs that must be carefully considered because they can affect performance and scalability differently.

✓ What does it mean to "normalize" a database table? And why would you do it? 1/1

When a database table is normalised, the columns and rows are arranged to reduce dependencies and redundancies. Usually, this is done to strengthen the database's integrity and structure.

You might want to normalise a database table for a number of reasons:

Reduced redundancy: Normalization aids in the elimination of redundant data, which can occupy extra storage space and make data updates more challenging.

Improved data integrity: By ensuring that data dependencies are correctly maintained, normalisation can help to lower the risk of data inconsistencies and errors.

Increased adaptability: Compared to non-normalized tables, normalised tables are typically more adaptable and simple to change.

Normalization may necessitate the introduction of additional tables and relationships, which can add to the complexity already present. Additionally, it may need the database to make extra joins in order to retrieve the data, which could affect performance.

✓ Suppose that you've been asked by a friend at a startup whether they should deploy a single database, a master-master pair of databases, a master database with one or more read-only replicas, or some other architecture altogether. What factors should inform your advice and guide their decision-making? 1/1

When giving a buddy advice on the ideal database architecture for their startup, you should take the following into account:

Data size: A single database may be adequate if the startup's data amount is anticipated to be small. However, a more scalable architecture, such as a master-master pair or a master with read-only copies, may be more suitable if the data size is anticipated to increase dramatically over time.

Data access patterns: Take into account the data access patterns. A master-master pair or a master with read-only replicas may be a better option if numerous users must access the data at once.

Data consistency: It's critical to take into account the level of data consistency needed. Strong consistency will be provided by a single database or a master-master pair, although this may come at the expense of decreased performance. Although it might be more efficient, a master with read-only clones might offer lower consistency.

Availability requirements for the data should also be taken into account. For applications that demand high availability, a single database or a master-master pair may be more appropriate because they can continue to process requests even if one of the servers fails.

Cost: It's important to take into account the costs associated with establishing and maintaining the various designs. The most economical choice might be a single database, but larger startups with more demanding needs might not be able to use it.

SQL

Download and install DB Browser from https://sqlitebrowser.org/ if you don't have it already.

Download IMDb.db from https://cdn.cs50.net/hls/2019/winter/lectures/7/imdb.db.

Open the latter with the former and use SQL (not, e.g., Google) to solve the problems below.

✓ Via which SQL query or queries might you determine how many actors have a first name of Meryl? 1/1

```
SELECT COUNT(*)
FROM people
WHERE name LIKE 'Meryl %';
```

✓ Roughly how many movies have been made about the Titanic? 1/1

51 ✓

✓ Via which SQL query or queries did you determine your answer? And why might your answer not be accurate? 1/1

```
SELECT COUNT(*)
FROM movies
WHERE title LIKE '%Titanic%';
```

This search will determine how many rows in the movies table have the word "Titanic" in the title column which may not be a possible 90's blockbuster most people would think of.

✓ In how many movies has Kevin Bacon acted? 1/1

57 ✓

✓ Via which SQL query or queries did you determine your answer? 1/1

```
SELECT COUNT(DISTINCT movie_id)
FROM cast_members
WHERE person_id IN (SELECT id FROM people WHERE name = 'Kevin Bacon');
```

Debrief

About how many MINUTES would you say you spent on this assignment in total? *

Just to set expectations for future students.

666