Introduction to Model Serving Infastructure

Congratulations! You passed!

Grade received 100% **Latest Submission Grade** 100% **To pass** 80% or higher

Go to next item

1. Why do models become more complex? **1 / 1 point**

- (•) To increase accuracy.
- ( ) To reduce GPU usage.
- ( ) To minimize latency.
- ( ) To cut down costs.

Correct
Absolutely! We apply more complex model architectures that allow including more features to increase accuracy.

2. What is the difference between optimizing and satisficing metrics? **1 / 1 point**

- (•) Optimizing metrics measure the model's predictive effectiveness while satisficing metrics specify operational constraints.
- ( ) Optimizing metrics assess model complexity while satisficing metrics evaluate operation costs.
- ( ) Optimizing metrics estimate the speed of the model's prediction latency while satisficing metrics deal with its precision.

Correct
Nailed it! First, aim to improve the model's predictive power until the infrastructure reaches a specific latency threshold. Then, assess the results to approve the model or continue working on it.

3. Which of the following are NoSQL solutions for implementing caching and feature lookup? (Select all that apply) **1 / 1 point**

- [ ] Amazon RDS
- [x] Google Cloud Memorystore

> **Correct**
> That's right! This database is a good choice for achieving sub-millisecond read latencies on a limited amount of quickly changing data retrieved by a few thousand clients.

- [x] Google Cloud Firestore

> **Correct**
> Right on! A good choice for millisecond read latencies on slowly changing data where storage scales automatically.

- [x] Amazon DynamoDB

> **Correct**
> Excellent! Amazon DynamoDB is a scalable low-read latency database with an in-memory cache.

4. True Or False: The main advantage of deploying a model in a large data center accessed by a remote call is that you can disregard costs in favor of model complexity. **1 / 1 point**

- ○ True
- ◉ False

> **Correct**
> Exactly! For example, Google constantly looks for ways to improve its resource utilization and reduce costs in its applications and data centers.

5. True Or False: As a rule, you should opt for on-device inference whenever possible. **1 / 1 point**

- ○ False
- ◉ True

> **Correct**
> Absolutely! Following this general rule enhances the user experience by reducing the app's response time. There are exceptions, though, such as medical diagnosis, in which the model must be as accurate as possible, and latency is not that important.