

✔ **Congratulations! You passed!**

Grade received **91.66%** To pass 80% or higher

Go to next item

Scaling Infrastructure

Total points 3

1. Why is managing scale paramount when serving a sophisticated model? (Select all that apply)

0.75 / 1 point

☒ The number of parameters increases considerably in more extensive and more sophisticated networks.

✔ **Correct**

Absolutely! The cost of training a network goes beyond the data—the more complex the network, the higher the number of parameters that need to be tuned and fine-tuned.

☒ The high volumes of requests to the model for inference can overwhelm the server.

✔ **Correct**

You've got it! When the model is deployed, too many inference requests can overwhelm the server, so the ability to scale the runtime inference and the training is vital.

☐ The costs of training deep neural networks with billions of operations on massive datasets are high.

☐ The changes in the dynamic environment in which the model operates may cause the model to drift.

You didn't select all the correct answers

2. True or False: The elasticity of vertical scaling allows to upgrade the hardware resources without taking the app offline.

1 / 1 point

☒ False

☐ True

✔ **Correct**

Indeed! When vertical scaling you must take the app offline as the upgrade involves or even replaces the entire hardware.

3. How do containers become lighter and more flexible than virtual machines?

1 / 1 point

☐ By having a hypervisor managing multiple instances of operating systems and apps.

☒ By using the same operating system for all partitions.

☐ By running the apps on a bin/library within an operating system that does not run on hardware.

✔ **Correct**

Nailed it! In containers, app instances don't require separate operating systems. In addition, the abstraction of having them run within a container runtime gives you greater flexibility.