Highlight Note

Module Review

In this module, you learned how to deploy a model to real-time and batch inference services.

Use the following review questions to check your learning.

Question 1

You want to deploy the model as a containerized real-time service with high scalability and token-based security.
What kind of deployment target should you use?

○ An Azure Container Instance (ACI)

○ An Azure Kubernetes Service (AKS) inference cluster

○ A multi-node compute cluster with GPUs

Check Answers

Question 2

Which functions must the entry script for a real-time service implement?

○ init and run

○ main and score

○ load and predict

Check Answers

Question 3

You want to implement a batch inference pipeline that distributes scoring on multiple nodes.
Which kind of pipeline step should you use?

○ PythonScriptStep

○ AdlaStep

○ ParallelRunStep

Check Answers