

PRACTICE QUESTIONS

Red Hat

EX267

Red Hat Certified Developer in AI

SHARED BY [u/ex267_student](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM Red Hat EX267, Red Hat Certified Developer in AI

Shared by u/ex267_student on aiexamcerts

10 practice questions for the Red Hat EX267 exam, Red Hat Certified Developer in AI, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/red-hat/ex267. Spotted a mistake? Email support@aiexamcerts.com with the exam code EX267 and the question number.

1. Which operator must be installed and healthy before worker nodes with NVIDIA GPUs can offer nvidia.com/gpu to pods?
 - A Triton Inference Server
 - B NVIDIA GPU Operator
 - C OpenShift Pipelines Operator
 - D Cluster Logging Operator

Answer: B

The NVIDIA GPU Operator installs the driver stack and device plugin that advertise GPUs to the scheduler. Triton Inference Server is a model server that uses a GPU but does not make GPUs schedulable, which is the tempting confusion. The OpenShift Pipelines Operator and Cluster Logging Operator are real cluster operators that provide no GPU resources.

2. The _____ resource is the custom resource used to turn OpenShift AI platform components on or off.

Answer: DataScienceCluster

The DataScienceCluster custom resource holds the managementState of each platform component. InferenceService is the tempting resource name for a serving task, but it describes a deployed model rather than the platform.

3. A model server pod stays Pending with an insufficient nvidia.com/gpu event while the GPU nodes are healthy. What is the most likely reason?
 - A The model format is not supported by the selected runtime, so the scheduler rejects the server pod
 - B Other pods already request all the available GPUs, so the scheduler cannot place this one
 - C The data connection secret holds expired credentials that the scheduler checks before it places the pod
 - D The route has no TLS certificate attached, which blocks the scheduler from placing the pod on a node

Answer: B

Pending with an insufficient GPU event means the scheduler found no node with free nvidia.com/gpu capacity. An unsupported model format would fail after the container starts, which makes it a tempting but wrong explanation. Expired credentials and certificates affect access, not scheduling, so the next step is to review GPU requests on the other pods.

4. A GPU workload can be scheduled on a cluster where the NVIDIA device plugin is absent.
- A True
 - B False

Answer: B (False)

A pod can request nvidia.com/gpu only when a device plugin advertises that resource to the scheduler, so without the plugin there is nothing to allocate. Node labels from Node Feature Discovery can exist without the plugin, which is the tempting partial setup. The statement is false.

5. Which two are valid ways to give a model deployment access to its data? (Choose two.)
- A Store the model in a workbench's ephemeral container filesystem
 - B Embed the model in a public Git repository only
 - C Reference a data connection that points to an S3 bucket
 - D Place the model files in a ConfigMap mounted into the pod
 - E Mount a PVC that holds the model files

Answer: C, E

A data connection lets the storage initializer download the model from S3 at startup, and a PVC can hold model files for the server to read. An ephemeral container filesystem is lost when the pod restarts, which makes it a tempting shortcut. Public repositories and ConfigMaps are not model storage the server can use.

6. Which storage keeps a workbench's notebook files after the workbench pod restarts?
- A An emptyDir volume inside the container, which is created with the pod and removed with it
 - B A persistent volume claim attached to the workbench
 - C The writable layer of the container filesystem that the runtime creates for each container
 - D A ConfigMap mounted at /tmp so that the notebook files are kept in cluster configuration

Answer: B

A PVC gives the workbench storage that outlives individual pod restarts. An emptyDir volume is deleted with its pod, and the container's writable layer is discarded when the container is recreated. A ConfigMap is not working storage, so mounting one at /tmp would not keep notebook files either.

7. A workbench shows ImagePullBackOff after it is started. What should the administrator check first?
- A Whether the kernel has enough memory to run the notebook once the container finally starts
 - B Whether the model's ServingRuntime lists the correct format for the model being deployed
 - C Whether the image reference is reachable and the pull secret is available to the namespace
 - D Whether the pipeline YAML compiles without warnings after the last change to the pipeline

Answer: C

ImagePullBackOff means the kubelet cannot pull the image, so the reference, registry access and pull secrets are the first checks. Kernel memory is the tempting next suspect, but it matters only after the container starts. ServingRuntimes and pipeline compilation belong to other workflows entirely.

8. Data connection credentials are commonly stored in a Kubernetes _____ object.

Answer: Secret

Data connections store S3 keys and endpoints in a Kubernetes Secret, which is the object meant for sensitive values. ConfigMap is the tempting alternative for configuration, but it is not used for credentials here.

9. Which serving approach hosts many smaller models in a shared set of model server pods?
- A Single-model serving with KServe in raw deployment mode
 - B Multi-model serving with ModelMesh
 - C A workbench that runs a TensorFlow kernel
 - D Distributed training with a Kubeflow training job

Answer: B

ModelMesh places several models into shared runtime pods and loads them on demand, which suits many small models. KServe single-model serving dedicates a model server to each model, which is the tempting alternative. Workbenches and training jobs do not serve inference traffic at all.

10. Which model server is widely used for serving large language models on OpenShift AI?
- A Jupyter Enterprise Gateway
 - B Apache Airflow
 - C vLLM
 - D Grafana Loki

Answer: C

vLLM is an inference engine for large language models and is offered as a serving runtime in OpenShift AI. Jupyter Enterprise Gateway launches notebook kernels, which makes it a tempting Jupyter-related choice. Airflow schedules workflows and Loki stores logs, so neither is a model server.