

PRACTICE QUESTIONS

Red Hat

EX267

Red Hat Certified Developer in AI

SHARED BY [u/studier_ex267](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM Red Hat EX267, Red Hat Certified Developer in AI

Shared by [u/studier_ex267](#) 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 command shows the management state of each OpenShift AI component on the cluster?

- A `oc get inferencesservice --all-namespaces`
- B `oc adm top nodes`
- C `oc get datasciencecluster -o yaml`
- D `oc get csv -n openshift-operators`

Answer: C

Reading the DataScienceCluster YAML shows the managementState of each component. Listing InferenceServices shows deployed models rather than platform components, and `oc adm top nodes` reports resource usage. Listing ClusterServiceVersions confirms the operator is installed but does not say which components are enabled.

2. Workbench notebook files are lost each time a workbench is stopped and started when a PVC is attached.

- A True
- B False

Answer: B (False)

A workbench with a PVC keeps its notebook files across stops and starts because the claim outlives the pod. Losing files is the common assumption about ephemeral containers, which is why it is tempting. Only deleting the PVC removes the stored data, so the statement is false.

3. Which label must a project namespace carry for the OpenShift AI dashboard to list it as a data science project?

- A `sidecar.istio.io/inject: 'true'`
- B `modelmesh-enabled: 'true'`
- C `app.kubernetes.io/managed-by: 'helm'`
- D `opendatahub.io/dashboard: 'true'`

Answer: D

The dashboard looks for the `opendatahub.io/dashboard` label on namespaces when it builds the project list. The Istio injection label controls proxy sidecars, and `modelmesh-enabled` marks a namespace for multi-model serving, so both are real labels that do not make a project appear. The `managed-by` label is a generic Kubernetes label with no dashboard meaning.

4. A data science project is backed by an OpenShift _____.

Answer: namespace

A data science project is a labeled namespace that scopes its workbenches, data connections and deployments. A cluster is far broader, which makes it the tempting but wrong answer. Namespace is the term the dashboard and RBAC use.

5. Which two are characteristics of KServe single-model serving? (Choose two.)

- A Serving resources are InferenceService and ServingRuntime objects
- B Multiple unrelated models share a single pod by design
- C It is limited to CPU inference only
- D Each deployed model gets its own model server
- E The workbench hosts the inference endpoint

Answer: A, D

Single-model serving dedicates a model server to each model, and it is built from InferenceService and ServingRuntime objects. Shared pods describe ModelMesh, which is the tempting confusion. CPU-only limits and workbench-hosted endpoints are not characteristics of KServe.

6. Which mechanism makes nvidia.com/gpu an allocatable resource on a worker node?

- A The kubelet's built-in CUDA driver, which it loads on every worker node during boot
- B Node Feature Discovery labels alone, which record detected hardware features on each node
- C The NVIDIA device plugin, deployed by the GPU Operator
- D The OpenShift router proxying GPU traffic from each notebook pod to the matching device

Answer: C

The device plugin registers GPUs with the kubelet, which advertises them as extended resources, and the GPU Operator deploys it. Node Feature Discovery only labels nodes with hardware features, so it is an incomplete answer that looks plausible. The kubelet does not carry a CUDA driver, and the router handles HTTP traffic rather than scheduling.

7. What is the role of Node Feature Discovery in a GPU-enabled OpenShift AI cluster?

- A It installs CUDA drivers into the kernel of each worker and keeps them updated over time
- B It schedules notebook pods onto the GPU nodes whenever a user starts a workbench there
- C It encrypts traffic between model servers and the clients that call their endpoints
- D It labels nodes with detected hardware features so GPU components can target them

Answer: D

Node Feature Discovery inspects hardware and applies node labels that other components use for targeting, including the GPU stack. Driver installation belongs to the GPU Operator, which makes the CUDA-driver option a tempting mix-up. Pod placement is the scheduler's job, and encryption has nothing to do with node labels.

8. ReadWriteOnce means a volume can be mounted read-write by a single node at a time.
- A True
 - B False

Answer: A (True)

ReadWriteOnce allows read-write mounting by one node at a time, while several pods on that node may still share it. It is often confused with ReadWriteMany, which allows many nodes. The statement is true.

9. How does a model deployment get a model file from an S3 bucket in OpenShift AI?
- A The S3 bucket is mounted as a hostPath volume on every node so the server reads it directly
 - B The model location references a data connection that the storage initializer uses at startup
 - C The model is baked into the workbench image during the notebook run and then pushed out
 - D The model URL points to a public website only, with no credentials, so it is fetched anonymously

Answer: B

A deployment that names a data connection lets the storage initializer fetch the model into the model server at startup. A hostPath mount would bypass the connection, which is the tempting workaround. Baking the model into a workbench image and relying on public URLs do not cover private buckets.

10. A model pod is Running, but inference requests return 503 errors. Which check is most useful first?
- A Delete the project namespace to reset the route and clear any stale state in the deployment
 - B Increase the storage size of the workbench volume so the model has room to load its weights from disk
 - C Recreate the data science pipeline that trained the model so the artifacts are regenerated
 - D Review the model server logs and readiness status to confirm the model loaded successfully

Answer: D

A Running pod can still be unready, and a 503 often means the model did not load, which the logs and readiness status reveal. Retraining the model is the tempting reaction, but the serving logs come first. Deleting the namespace and enlarging a workbench volume do not address a model server that has not become ready.