KaozhengPro

IT認證考試題庫

專業平臺

考證寶提供最新考古題與模擬試題
協助您高效通過認證考試

www.kaozhengpro.com

Exam : **NS0-901**

Title : NetApp Certified AI Expert Exam

Version : DEMO

1.An organization is developing a new AI-powered application. The initial phase involves feeding a curated 50 TB dataset of labeled images into a complex neural network, allowing the model to learn and adjust its internal parameters over millions of iterations. The second phase involves deploying this finalized model to a web service where it will process single, user-uploaded images and return a classification in real-time.
Which statement accurately describes these two phases?
A. Phase 1 is inferencing, and Phase 2 is training.
B. Phase 1 is training, and Phase 2 is inferencing.
C. Both Phase 1 and Phase 2 are examples of training.
D. Both Phase 1 and Phase 2 are examples of inferencing.
Answer: B

2.An AI architect is planning the resource allocation for a new project. The primary task is to process millions of unlabeled customer reviews to identify naturally occurring groups or themes without any prior guidance.
The project requirements are summarized below:
Task: Discover hidden patterns in text data
Input_Data: 10 million unlabeled text reviews
Output: Clustered groups of related reviews
Supervision: None
Which type of machine learning algorithm is required for this task?
A. Supervised learning
B. Reinforcement learning
C. Unsupervised learning
D. Predictive learning
Answer: C

3.A financial services company has deployed a real-time fraud detection model at the edge. The model is designed for low-latency inference. However, monitoring reports indicate that the infrastructure costs are excessively high, and GPU utilization is consistently low. The architect reviews the deployment configuration.
Instance_Type: NVIDIA DGX A100 (8 GPUs)
Storage_Tier: High-Performance All-Flash (NetApp ASA)
Network: 100GbE RoCE
GPU_Utilization_Avg: 5%
Monthly_Cost: $15,000
Workload_Profile: Low-volume, sporadic, real-time predictions
What is the most likely cause of the high costs and low utilization?
A. The network latency is too high for an edge deployment.
B. The storage tier is too slow, causing the GPUs to wait for data.
C. The compute and storage infrastructure is sized for a large-scale training workload, not a lightweight inference workload.
D. The model was trained using supervised learning, which is inefficient for fraud detection.
Answer: C

4.A research institute is designing an infrastructure to support its entire AI drug discovery pipeline. The pipeline has two distinct workload requirements:
1. Training: A team of data scientists needs to train several large transformer models simultaneously using a 500 TB dataset of genomic sequences. This process requires maximum data throughput to keep the GPUs saturated.
2. Inference: Once trained, the models are deployed to an internal web portal where researchers submit individual protein sequences for analysis. These queries must return results with the lowest possible latency.
Which infrastructure design best satisfies both requirements? (Choose 2.)
A. Deploy a large NetApp ASA cluster with GPUDirect Storage enabled for the training environment.
B. Use NetApp StorageGRID as the primary storage for both the training and low-latency inference workloads.
C. Implement NetApp FlexCache on smaller nodes at the network edge to serve the inference requests.
D. Use a single, large Cloud Volumes ONTAP instance in a public cloud to handle both workloads to simplify management.
E. Configure QoS minimums on the training volumes to ensure they do not impact inference performance.
Answer: A, C

5.An online retail company's recommendation engine, which provides real-time product suggestions to users, is experiencing unacceptable latency. The inference application is running on a correctly-sized edge server, but user requests are taking over 500ms to process. An architect reviews the data access pattern and infrastructure diagram.

Application_Location: Edge Server (In-store)
Data_Source_Location: Core Data Center (On-premises ONTAP)
Data_Required_for_Inference: User profile data, product catalog vectors
Network_Path: Edge -> WAN -> Core Data Center
Observed_Latency: 550ms

What is the most likely cause of the high inference latency?
A. The edge server has insufficient CPU resources to run the model.
B. The on-premises ONTAP system is not configured for high-throughput.
C. The model is too large to fit into the edge server's memory.
D. Every inference request requires a high-latency round trip over the WAN to fetch data from the core data center.
Answer: D