AI/ML Deployment

Product Code: CNS-INF-A-SVC-DEP-GPT

At-a-Glance

Phase: Deploy

The Nutanix Artificial Intelligence / Machine Learning (AI/ML) Deployment service accelerates the deployment of a Nutanix-provided bootstrap AI/ML inference solution based on a Generative Pre-trained Transformer (GPT) workload on Nutanix Cloud Infrastructure (NCI). The deployed solution will validate that the NCI cluster can be used for virtual machine (VM)- or container-based inference. It also provides an API endpoint for the customer to build out the AI/ML application. This offer is ideal for the Deploy stage of the Nutanix GPT-in-a-box Solution journey.

Service Scope

Highly skilled consultants deploy the Nutanix-provided bootstrap solution by enabling NCM Self-Service in Prism Central, deploying a VM-based blueprint from the NCM Self-Service Marketplace, and configuring it. After the deployment of the solution, the consultant will demonstrate the model validation with sample data.

This service includes the following activities:

- Review the NCI cluster configuration that runs the AI workload
- Deploy Nutanix-provided bootstrap application that will provide a REST API endpoint for the application integration
  - Configure NCM Self-Service
  - Enable NCM Self-Service in Prism Central
    - Deploy one Ubuntu blueprint from the NCM Self-Service Marketplace as the AI compute instance, which includes the following:
      - Configure the compute instance with GPU passthrough
      - Install NVIDIA drivers to use the GPU
      - Configure the compute instance with shared storage (NUS Files or Objects) to store a large language model (LLM)
      - Install the LLM prerequisites, including the Nutanix LLM package and Python libraries
  - Download model files from the LLM provider and generate the model archive
  - Deploy PyTorch serving framework to run the LLM
- Demonstrate the bootstrap application with sample data

Limitations

- Excludes training a new LLM
- Excludes creation or updates to existing design documentation

- Excludes NCI cluster, NUS, and Prism Central deployment.

  Note: *NCI Cluster Deployment or Expansion Pro Edition* is recommended

- For each quantity purchased, deployment is limited to 1 on-premises NCI cluster
- NCM Self-Service configuration is limited to 1 blueprint deployment from the NCM Self-Service Marketplace

Supported Hypervisors

- Nutanix AHV

Prerequisites

- Fully supported and functional on-premises NCI cluster that meets all product requirements for GPT-in-a-Box, NUS, Prism Central, NCM Self-Service, and a supported GPU

  Note: For information on the requirements for NCI Clusters, see Field Installation Overview in the *Field Installation Guide* on the Nutanix Support Portal.

  Note: For information on the requirements for deploying NCM Self-Service, see *Calm Prerequisites and Deployment in Calm Administration and Operations Guide* on the Nutanix Support Portal.

  Note: For information on NUS Files Prerequisites, see Prerequisites in *Nutanix Files User's Guide* on the Nutanix Support Portal.

- Completed Pre-Install Questionnaire

Required Product Licenses

- Nutanix Cloud Infrastructure (NCI) Ultimate Edition
- Nutanix Unified Storage (NUS) Pro Edition

Deliverables

- Project Kickoff
- Project Schedule
- Project Status Report(s)
- Deployment of bootstrap solution
- Usable API endpoint
- Project Closeout

Duration

Typically up to 2 days

Related Products

- Nutanix Cloud Infrastructure (NCI)
- Nutanix Unified Storage (NUS)
- Nutanix Cloud Manager (NCM)

Terms and Conditions

This document contains the entire scope of the service offer. Anything not explicitly included above is out of scope. This service offer is subject to the Nutanix Services General Terms and Conditions that can be viewed at https://www.nutanix.com/support-services/consulting-services/terms-and-conditions

Tel 1 (855) 688-2649 · **Fax** 1 (408) 916-4039 · **Email** info@nutanix.com ·