VMware Private AI Foundation with NVIDIA 5.2.x Release Notes

VMware Cloud Foundation 5.2

vmware by Broadcom

You can find the most up-to-date technical documentation on the VMware by Broadcom website at:

https://docs.vmware.com/

VMware by Broadcom
3401 Hillview Ave.
Palo Alto, CA 94304
www.vmware.com

Contents

Introduction

1

You can use VMware Private AI Foundation with NVIDIA to run generative AI workloads by using accelerated computing from NVIDIA, and virtual infrastructure management and cloud management from VMware Cloud Foundation.

VMware Private AI Foundation with NVIDIA is based on VMware Cloud Foundation, with added GPU- and AI-centric functionality in VMware Aria Operations, VMware Aria Automation, and VMware Data Services Manager.

VMware Private AI Foundation with NVIDIA 5.2.1 | 09 OCT 2024
VMware Private AI Foundation with NVIDIA 5.2 | 23 JUL 2024
Check for additions and updates to these release notes.

What's New

2

Read the following topics next:

VMware Cloud Foundation 5.2.1

- The vSphere Client provides a guided deployment workflow that you can follow to set up a GPU-enabled workload domain with a Supervisor instance for running AI workloads and VMware Data Services Manager.
- The self-service catalog of VMware Aria Automation provides separate items for Retrieval-augmented generation (RAG) AI workloads with pgvector databases or for standalone pgvector databases on top of VMware Data Services Manager.
- The AI self-service catalog items in VMware Aria Automation support collection and exposure of GPU-related metrics from deployed AI workloads by using Data Center GPU Manager (DCGM) Exporter to Prometheus and Grafana.
- You can turn on JupyterLab authentication when deploying a deep learning VM with PyTorch or TensorFlow by using the self-service catalog in VMware Aria Automation.
- You can store ML models in a central Harbor registry, accessible from a deep learning VM, for validation, governance, and portability across organizations and environments.

VMware Cloud Foundation 5.2

- The self-service catalog of VMware Aria Automation provides separate items for provisioning RAG workloads.
- You can add integration with VMware Data Services Manager to the self-service catalog of VMware Aria Automation.

VMware Deep Learning VM Image

See VMware Deep Learning VM Image Release Notes.

Compatibility

3

The functionality of VMware Private AI Foundation with NVIDIA is available in VMware Cloud Foundation and certain versions of VMware Aria Operations, VMware Aria Automation, and VMware Data Services Manager running in VMware Cloud Foundation.

VMware Cloud Foundation Version	Versions of VMware Aria Components and VMware Data Services Manager
VMware Cloud Foundation 5.2.1	■ VMware Aria Operations 8.18.1 ■ VMware Aria Automation 8.18.1 ■ VMware Data Services Manager 2.1
VMware Cloud Foundation 5.2	■ VMware Aria Operations 8.18 ■ VMware Aria Automation 8.18 ■ VMware Data Services Manager 2.1

Installation and Upgrade

4

VMware Private AI Foundation with NVIDIA runs on top of VMware Cloud Foundation and is based on the components included in the bill of materials of VMware Cloud Foundation.

For information about deploying VMware Cloud Foundation, see VMware Cloud Foundation Deployment Guide. For information about upgrading VMware Cloud Foundation, see VMware Cloud Foundation Lifecycle Management.

For information about adding the functionality of VMware Private AI Foundation with NVIDIA on top of VMware Cloud Foundation, see Preparing VMware Cloud Foundation for Private AI Workload Deployment.

License Information

5

VMware Private AI Foundation with NVIDIA is available under a solution license for VMware Cloud Foundation. You also need an NVIDIA AI Enterprise license for the host driver VIB file for ESXi hosts and the guest OS drivers, and for downloading AI container images from the NVIDIA NGC catalog. See Requirements for Deploying VMware Private AI Foundation with NVIDIA.

Documentation

6

Examine the VMware Private AI Foundation with NVIDIA Guide for an overview and how-to instructions for running AI workload and storing ML models in a VMware Cloud Foundation environment.

Known Issues for VMware Cloud Foundation 5.2.1

7

- **Tasks in SDDC Manager that are synced with vCenter Server are reported as failed in the vSphere Client but as in-progress in SDDC Manager**

  If an operation in SDDC Manager, started from the vSphere Client, such as creating a workload domain in the Private AI Foundation guided deployment, completes with error, you see the task as failed in the vSphere Client while the SDDC Manager UI still shows it as in-progress.

  Take into account the task status in the vSphere Client.