TECHNICAL WHITE PAPER: **July 2024**

VMware Cloud Foundation with IBM watsonx AI Platform

VMware Validated Solution

vmware by Broadcom

Table of contents

Note: This solution provides general solution design and deployment guidelines for running the IBM watsonx platform on VMware Cloud Foundation. The hardware used in this paper is not strictly tied to the solution architecture. The solution architecture applies to any compatible hardware platforms running VMware Cloud Foundation.

Executive Summary

Business Case

VMware Private AI™ is an architectural approach that aims to balance the business gains from AI with the privacy and compliance needs of the organization. VMware Private AI focuses on the platform and infrastructure architecture built-in support of AI, which can be deployed in public clouds, virtual private clouds, data centers, and edge locations.

VMware and IBM have developed a validated solution architecture for Private AI that will allow mutual customers to leverage secure, compliant, and trusted hybrid cloud and on-prem environments for the training and fine-tuning of ML models – including Large Language Models (LLMs) for generative AI use cases. For example, moving data from a bank or financial institution's data center in one country to a hyperscaler may be prohibited due to regulatory restrictions or the data is too large, and attempting to transfer it from one location to another would be impossible.

By having the choice of when, where, and how to integrate generative AI technologies with VMware Cloud Foundation, enterprises are empowered to quickly train and deploy custom AI capabilities across their enterprise while retaining full control and compliance over their data. They also gain the best innovations from VMware's software offerings in a unified stack to deliver a consistent environment integrated with the data and AI capabilities brought by the IBM watsonx AI platform.

In this solution, we provide the well-architected solution design guidelines for running IBM watsonx on VMware Cloud Foundation.

Virtualization Brings Operational Savings to Containerized Workloads

With a well-integrated Kubernetes platform such as OpenShift or Tanzu, vSphere provides a developer-ready infrastructure for container platforms including OpenShift. Well-known vCenter tools and processes can now manage both traditional VM-based and containerized workloads across your hybrid cloud. vSphere brings trusted capabilities such as VMware vSphere High Availability (vSphere HA) and policy-based management to ensure availability and resiliency for all workloads. vSphere also enhances the security of containers by naturally providing isolation of the VMs. In addition, vSphere life-cycle management and enterprise resiliency reduce the administration time required to manage bare metal updates and failures. All the above benefits help improve daily operational efficiency for DevOps.

To learn more about these benefits in detail, check out Why Choose VMware Virtualization for Kubernetes and Containers.

Audience

This VMware Validated Solution paper is intended for the following audiences:

- Corporate CTOs and CIOs who are architecting any flavors of Kubernetes in a private data center or in a hybrid cloud.
- vSphere VI administrators who are familiar with VMware virtualized infrastructure and need to deploy and manage OpenShift in a virtualized environment.
- DevOps who are deploying, managing, or using OpenShift on vSphere now it is VCF.
- Enterprises and users who want to run AI applications.
- Any other engineer/operator/end-user who are interested in OpenShift/Kubernetes/vSphere and have the basic knowledges about VMware Cloud Foundation, VSAN, NSX, Cloud Native Storage (CNS), Container Storage Interface (CSI), OpenShift, Kubernetes, and AI.

Technology Overview

Solution technology components are listed below:

- VMware Stack
    - VMware Cloud Foundation
    - Kubernetes vSphere CSI VMware Container Networking™ with Antrea™ for OpenShift
    - VMware NSX Container Plug-in for OpenShift
- IBM Stack

  - IBM Cloud Satellite
  - IBM watsonx Platform
  - IBM Cloud Pak for Data
  - Red Hat OpenShift Container Platform
  - IBM OpenShift Data Storage
- Nvidia
  - Nvidia GPUs (NVAIE)

VMware Cloud Foundation

VMware Cloud Foundation is an integrated software stack that combines compute virtualization (VMware vSphere), storage virtualization (VMware vSAN), network virtualization (VMware NSX), and cloud management and monitoring (VMware vRealize® Suite) into a single platform that can be deployed on-premises as a private cloud or run as a service within a public cloud. This documentation focuses on the private cloud use case. VMware Cloud Foundation bridges the traditional administrative silos in data centers, merging compute, storage, network provisioning, and cloud management to facilitate end-to-end support for application deployment. See Getting Started with VMware Cloud Foundation for details.

NVIDIA GPUs

Nvidia A100 GPUs, part of the Nvidia Ampere GPU architecture, are powerful graphics processing units designed for high-performance computing (HPC), artificial intelligence (AI), and data analytics workloads. These GPUs are built on a 7nm process and feature several key specifications and technologies:

1. Architecture: The A100 GPUs are based on the Nvidia Ampere architecture, which introduces improvements in performance, energy efficiency, and support for AI workloads.
2. Tensor Cores: A notable feature is the inclusion of Tensor Cores, specialized hardware designed to accelerate AI and deep learning workloads by efficiently performing matrix multiplication operations.
3. CUDA Cores: The A100 GPUs have a large number of CUDA cores, providing parallel processing capabilities for a wide range of computational tasks.
4. High Memory Capacity: A100 GPUs are equipped with high-bandwidth memory (HBM2) with a large memory capacity, allowing them to handle large datasets efficiently.
5. NVLink: Nvidia A100 GPUs support NVLink, a high-speed interconnect technology that enables multiple GPUs to communicate with each other at high speeds, enhancing scalability for multi-GPU configurations.
6. Multi-Instance GPU (MIG): A feature that allows a single A100 GPU to be partitioned into multiple instances, each capable of running different workloads simultaneously.
7. NvLink Connection: A100 GPUs can be connected using Nvidia's NVLink technology, enabling fast communication between GPUs in multi-GPU setups.

These GPUs are widely used in data centers and supercomputing environments for applications such as scientific simulations, deep learning training, and inference tasks due to their high computational capabilities.

Red Hat OpenShift Container Platform

Red Hat OpenShift Container Platform ships with Red Hat Enterprise Linux® CoreOS for the Kubernetes control plane nodes and supports both Red Hat Enterprise Linux CoreOS and Red Hat Enterprise Linux for worker nodes.

See https://www.redhat.com/en/technologies/cloud-computing/openshift for detailed information regarding the OpenShift Container Platform.

IBM Cloud Satellite

IBM Cloud Satellite® provides a distributed cloud architecture that brings the scalability and flexibility of public cloud services to the applications and data that run in your secure private cloud. With IBM Cloud Satellite®, you can use your own compute infrastructure that is in your on-premises data center, other cloud providers, or edge networks to create a Satellite location. Then, you can use the capabilities of Satellite to run IBM Cloud services on your infrastructure, and consistently deploy, manage, and control your app workloads. From a single pane of glass, you can manage workloads that run across the infrastructure from your Satellite locations.

IBM watsonx AI and Data Platform

IBM watsonx AI and data platform include three core components and a set of AI assistants designed to deploy and embed AI with trusted data across the business for generative AI as well as traditional AI use cases.

watsonx.ai - Allows you to build a studio for foundation models, generative AI, and machine learning. With watsonx.ai, you can train, validate, tune, and deploy foundation and machine learning models with ease.

watsonx.data - Scale analytics and AI workloads for all your data, anywhere with watsonx.data, the industry's only data store that is open, hybrid and governed.

watsonx.governance - Accelerate responsibility, transparency and explainability in your data and AI workflows with watsonx.governance. This solution helps you direct, manage, and monitor your organization's AI activities.

See https://ibm.com/watsonx for detailed information regarding the watsonx platform.

IBM Cloud Pak for Data

IBM Cloud Pak® for Data is a cloud-native solution that enables you to put your data to work quickly and efficiently.

Cloud Pak for Data fosters productivity by enabling users to find existing data or to request access to data. With modern tools that facilitate analytics and remove barriers to collaboration, users can spend less time finding data and more time using it effectively.

The Cloud Pak for Data can run on your Red Hat® OpenShift® cluster, whether it's behind your firewall or on the cloud.

- **On the cloud** - If you have an OpenShift deployment on IBM® Cloud, AWS, Microsoft Azure, or Google Cloud, you can deploy Cloud Pak for Data on your cluster.
- **On-premises** - Prefer to keep your deployment behind a firewall? You can run Cloud Pak for Data on your private, on-premises cluster.

Solution Configuration

This section introduces the following resources and configurations:

- Well-architected Solution
    - Hardware resources
    - Software resources
        - VMware Cloud Foundation Installation
        - Red Hat OpenShift installation with Antrea
    - IBM Satellite Deployment
    - IBM watsonx Deployment
        - OpenShift cluster deployment via Cloud Satellite
        - IBM Cloud Pak for Data installation

- IBM watsonx.ai installation
- IBM watsonx.data installation
- IBM watsonx.governance installation
- IBM Foundation Models installation

Solution Architecture

In this solution, the VMware Cloud Foundation test environment was composed of a management domain and a workload domain.

In this solution, the VMware Cloud Foundation validation environment was composed of a standard architecture, including a management domain and a GPU-enabled workload domain.

Different workload domains can serve different business purposes. The one-to-many mapping simplifies the overall management of the whole VMware Cloud Foundation environment. We used Nvidia GPUs for the workload domain to validate the IBM watsonx stack.

We deployed the OpenShift Container Platform in one of the GPU-enabled workload domains. The workload domain also contains the VMware NSX® Edge™ nodes. All other infrastructure VMs were placed in the shared management workload domain (Figure 1).

This figure only shows the workload domain of OpenShift Container Platform that we focused on in this architecture for deploying the IBM watsonx stack.

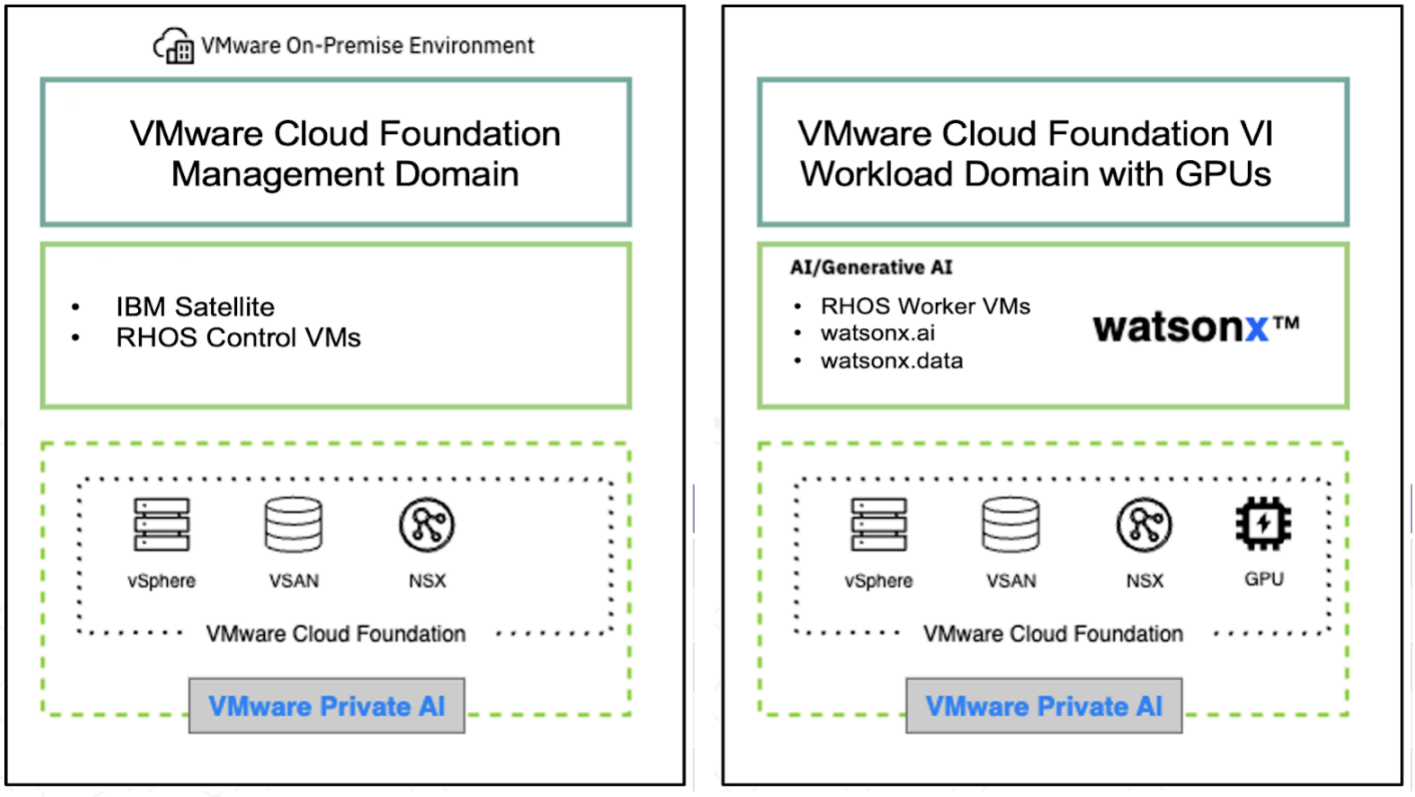

Figure 1 – On-premises deployment of watsonx stack

Notation in Figure 1:

- IBM Cloud Satellite: This is the optional component to deploy OpenShift Container Platform based on the automation.
- Control VMs: These are the control plane nodes of Kubernetes deployed in the management domain and managed by OpenShift.

- Worker VMs: These are the worker nodes of Kubernetes deployed and managed by OpenShift. We deployed 6 worker nodes in the VI workload domain enabled with GPUs to get ready for the IBM watsonx components mentioned below.

The above architecture is called VMware Cloud Foundation's standard architecture.

VMware Cloud Foundation Standard Architecture

For the standard architecture, we created a 4-node ESXi cluster for the VMware Cloud Foundation management domain, running management virtual machines and appliances. The management domain can be shared with other workload domains.

Table 1. Management Domain VMs

VM Role	vCPU	Memory (GB)	VM Count
Management Domain vCenter Server	4	16	1
SDDC Manager	4	16	1
Management Domain NSX Manager	6	24	3
Workload Domain NSX Manager	12	48	3
Workload Domain vCenter Server	8	28	1

For the workload domain, we created another 4-node ESXi cluster with a separate NSX Fabric, deployed an NSX Edge Cluster, and deployed the OpenShift RHOS8 VMs in the workload domain using IBM Cloud Satellite.

Table 2 shows the deployment of the workload domain edge nodes and OpenShift VMs. For the workload domain edge node, we recommend that NSX Edge transport nodes are deployed with "Large" form factor.

Table 2. Workload Domain VMs

VM Role	Minimum vCPU	Minimum Memory (GB)	Storage	Deployment Size	VM Count
Workload Domain Edge node	8	32	200 GB	Large	6

OpenShift Control Plane Nodes	16	64	120 GB for OS	n/a	6
OpenShift Compute (Worker) Nodes	16	64	120 GB for OS and 800 GB RAW	n/a	6

Hardware Resources

In this solution, for the workload domain of OpenShift, we used a total of four Dell R750 nodes.

Each ESXi node in the cluster had the following configuration, as shown in Table 3.

Table 3. Hardware Configuration for Workload Cluster

PROPERTY	SPECIFICATION
Server model name	Dell PowerEdge R750
CPU	2 x Intel Xeon Gold 6338 @ 2.00GHz, 64-Core Processor
RAM	512 GB
Network adapter	2 x Broadcom Corporation NetXtreme BCM5720 Gigabit Ethernet 2 x 25Gbits/s Broadcom NetXtreme E-Series Advanced Dual-port 25Gb OCP 3.0 Adapter
Storage adapter	1 x Dell HBA355i Front Adapter & Dell Ent NVMe CM6 MU 1.6TB
Disks	Cache - 2 x 800GB Write-intensive SAS SSDs Capacity - 6 x 1.46TB Read-intensive SAS SSDs
GPU	NVIDIA A100/40GB x 2

Software Resources

Table 4 shows the software resources used in this solution.

Table 4. Software Resources

Software	Version	Purpose
VMware Cloud Foundation	5.1 Build 22688368	A unified SDDC platform that brings together VMware vSphere, vSAN, NSX, and optionally, vRealize Suite components, into a natively integrated stack to deliver enterprise-ready cloud infrastructure for the private and public cloud. See VMware Cloud Foundation 5.1 Release Notes for details. We used VMware Cloud Foundation 5.1 in this solution, a later minor version is also supported. See Preparing to install on vSphere for VMware vSphere infrastructure requirements.

NVAIE/GPU Driver Version		
Red Hat OpenShift	4.12	The version of OpenShift software being tested in this solution. See OpenShift Container Platform 4.12 Release Notes for the latest release introduction.
IBM watsonx.ai	4.8.2	Build with our new studio for foundation models, generative AI and machine learning. Train, validate, tune, and deploy foundation and machine learning models with ease.
IBM watsonx.data	4.8.2	Scale AI workloads, for all your data, anywhere with a fit-for-purpose data store built on an open lakehouse architecture.
IBM Cloud Pak for Data	4.8.2	A fully integrated data and AI platform that modernizes how businesses collect, organize, and analyze data to infuse AI
IBM Satellite	N/A	IBM Cloud Satellite is a secure cloud platform that allows you to use managed services anywhere, easily connect to IBM Cloud, and seamlessly deploy to multiple clusters.

Solution Deployment Workflow

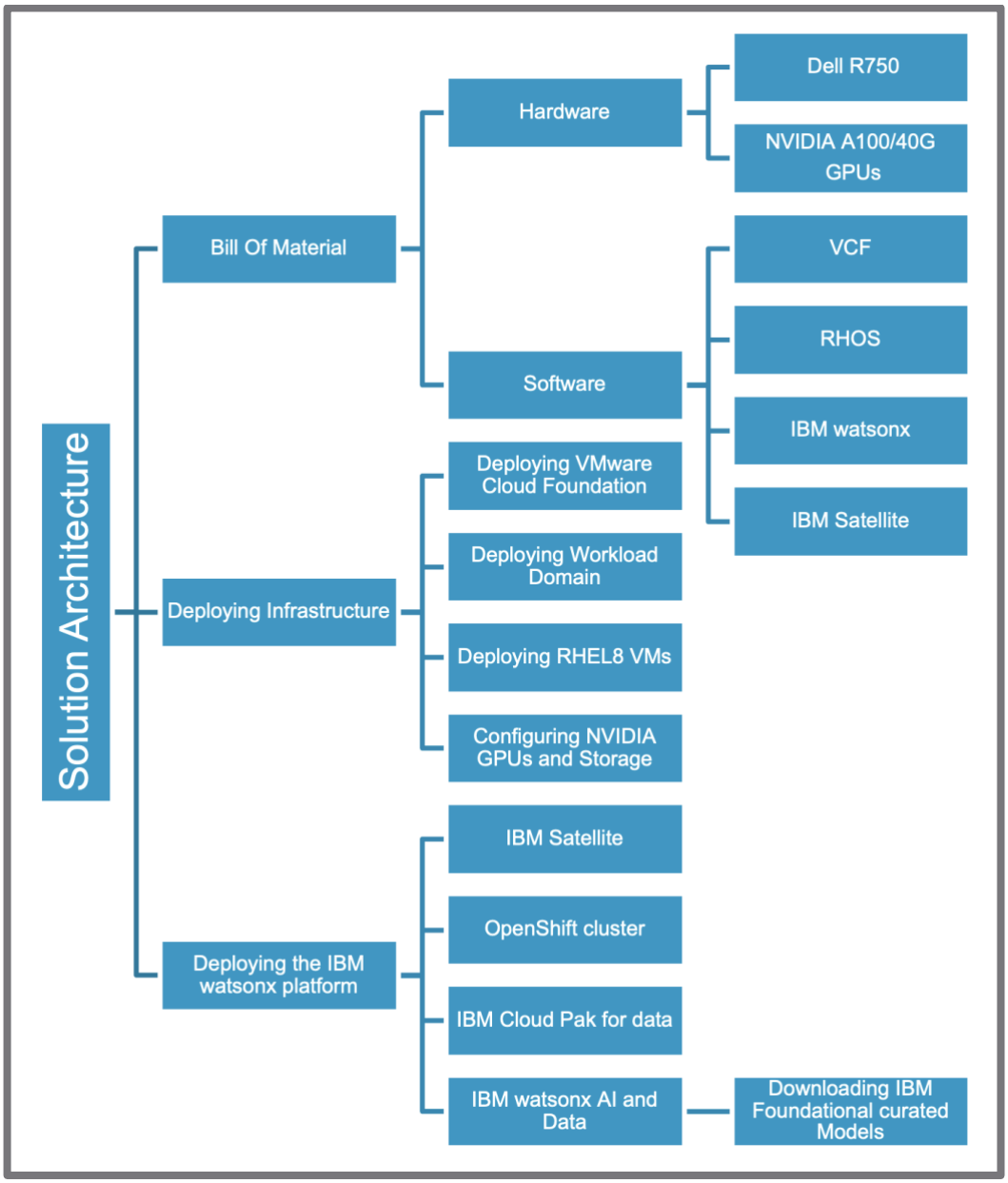

Figure 2 – Solution Architecture infrastructure deployment

VMware Cloud Foundation Installation

Follow the Official Document for detailed information about VMware Cloud Foundation installation steps.

The key steps for VMware Cloud Foundation installation are as follows:

1. Deploy a management domain.
2. Add ESXi hosts into the system.
3. Create a workload domain with the idle ESXi hosts.

Perform the VMware Cloud Foundation bring-up process to complete the management domain and commission the 3 hosts to the inventory to create the VI workload domain. The workload domain will have 3 hosts each with 2 NVIDIA A100/40G GPUs.

After the installation, just make sure that NSX and vSAN are successfully enabled on this workload domain. NSX and vSAN are integrated into this solution and will be used in the successive configurations of OpenShift.

RHEL VM Installation

To deploy the Red Hat OpenShift cluster using IBM Satellite, we need to first deploy the RHEL version 8 virtual machines by leveraging the standard process of creating a VM. Each RHEL virtual machine as either a control plane node or worker node was sized 16 vCPUs and 64GB memory and required storage for OS and RAW file system of 800 GB as one of the key requisites of the IBM watsonx.ai deployment. For detailed instructions, follow the documentation.

Pre-requisites are well-defined steps can be found here & the GPU enablement steps are here.

IBM Cloud Satellite Installation

The infrastructure typically on-premises data centers, other cloud providers, or edge networks & defines the location by attaching hosts (compute sources) from your chosen infrastructure, for us these are RHEL VMs. These VMs will form the basis of your Satellite environment. The Satellite connector provides a way to secure communications from IBM Cloud to the on-premises resources via a connector agent. Upon successful deployment, you can leverage various IBM Cloud services like deploying a Red Hat OpenShift cluster on the infrastructure hosts aka RHOS VMs within your Satellite location.

Remember that Satellite allows you to deploy apps and run them anywhere, leveraging your own infrastructure while benefiting from the capabilities of IBM Cloud services. For more detailed instructions, refer to the IBM Cloud Satellite documentation.

Red Hat OpenShift Installation

There are five methods to install OpenShift on vSphere: Agent-based Installer, Assisted Installer, 'User Provisioned Infrastructure' (UPI), Installer-Provisioned Infrastructure (IPI), and IBM Cloud Satellite. The IBM Cloud Satellite is the easiest way of deploying as it has more automation during deployment and follow-up management.

As of OpenShift version 4.12, which is used in this solution architecture, Satellite is fully supported and compatible across the VMware Cloud Foundation BOM for installing OpenShift on VMware Cloud Foundation including Satellite Storage templates.

During the installation process, we mainly used the IBM Satellite documentation and OpenShift on VMware Cloud Foundation for reference.

We will demonstrate the major steps and highlights in this solution architecture while the above documents have more detailed information about the explanation of each step.

IBM Cloud Pak for Data Installation

You can install one or more instances of IBM Cloud Pak for Data on your Red Hat OpenShift Container Platform cluster. Each instance of Cloud Pak for Data has its own project for operators and its project for the Cloud Pak for Data control plane and services (also called operands). The details can be found here.

The Cloud Pak for Data installation is broken up into the following phases:

1. Setting up a client workstation
2. Setting up a cluster
3. Collecting required information

IBM watsonx Configuration

VMware Private AI with IBM provides enterprise clients the flexibility to enable generative AI use cases at all levels of business, no matter where their mission-critical operations reside. The combination of VMware Cloud Foundation and Red Hat OpenShift enables enterprises to access IBM watsonx in on-premises or private cloud environments as well as hybrid cloud with watsonx SaaS offerings on IBM Cloud. Enterprise clients can utilize IBM watsonx AI capabilities, data management, and governance across public clouds, virtual private clouds, data centers, and edge sites, helping to enable responsible, transparent, and explainable AI models.

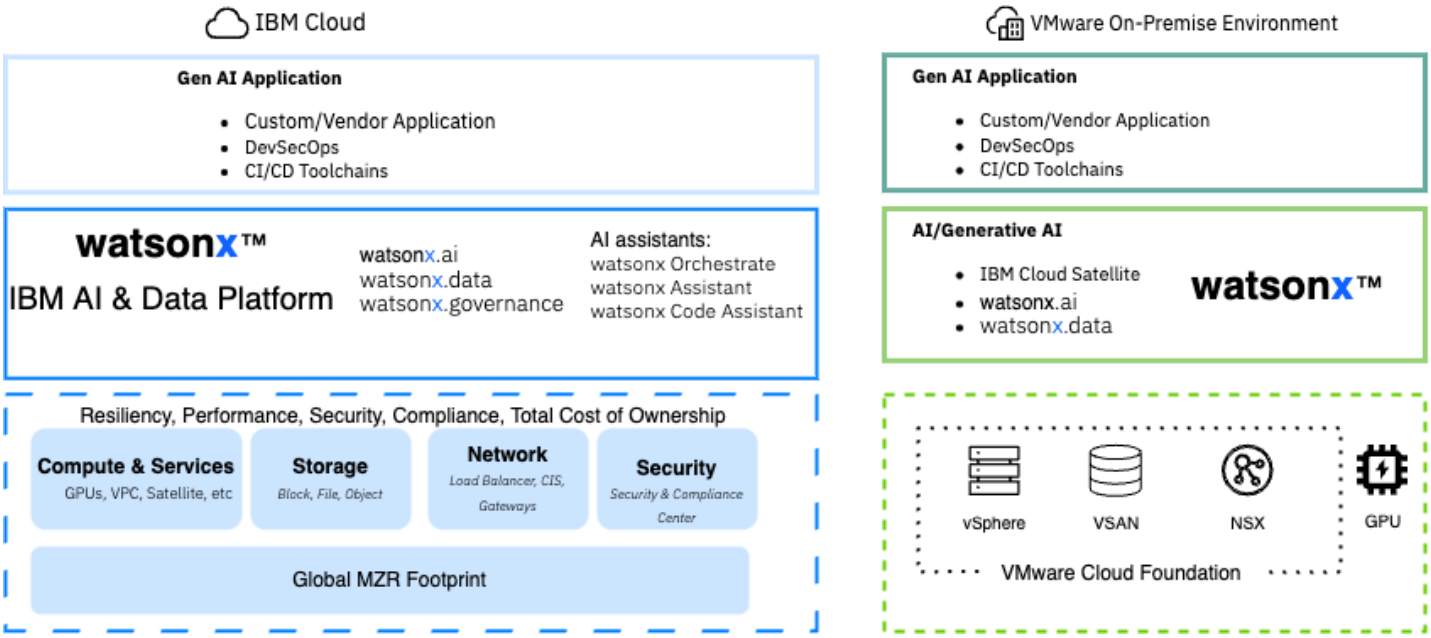

Figure 3 – High-level Architecture of on-premises VCF with IBM watsonx

As shown in Figure 1 above, VMware Private AI is available across hybrid multi-cloud deployments. Using Red Hat OpenShift, enterprise clients can now bring watsonx on-prem for VMware-deployed workloads.

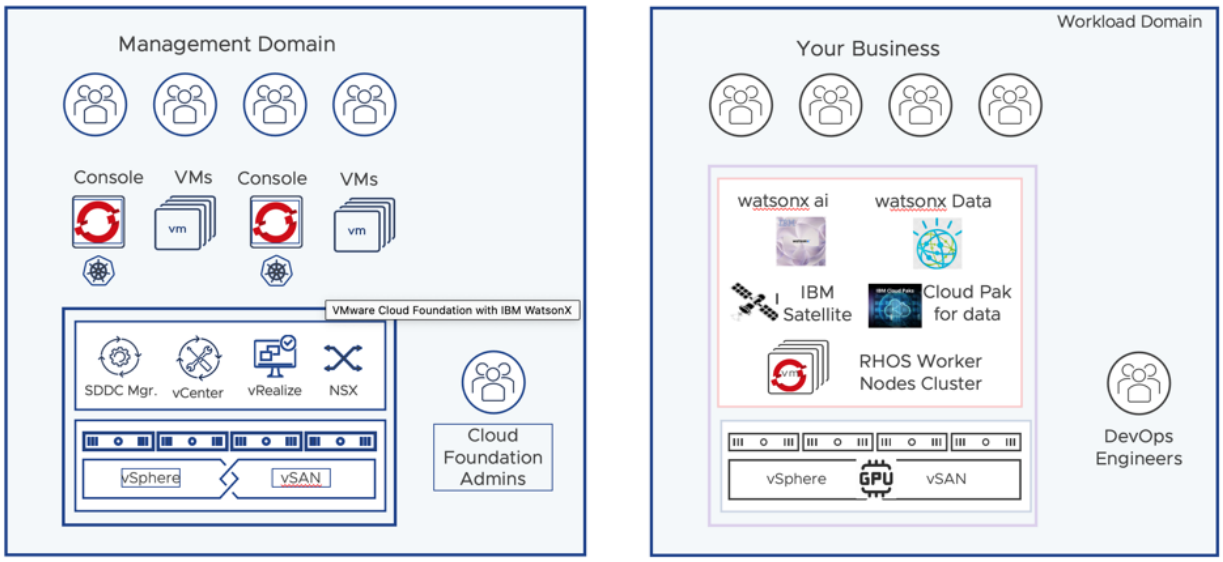

Figure 4 – Detailed Architecture of on-premises VCF with IBM watsonx

As mentioned above, the watsonx platform comprises three key components, allowing you to customize your AI solution:

watsonx.ai - Allows you to build with a studio for foundation models, generative AI and machine learning.

watsonx.data - Scale analytics and AI workloads for all your data anywhere with the industry's only data store that is open, hybrid and governed.

watson.governance - Accelerate responsibility, transparency and explain ability in your data and AI workflows.

Access IBM watsonx in private environments

VMware Private AI enables enterprises to access IBM watsonx on-premises and in the IBM cloud for the secure training and fine-tuning of their models with the watsonx platform. It provides training ML models/LLMs using proprietary data, especially within regulated industries. The private cloud option can help organizations reduce the amount of capital expenditure needed for training GenAI models.

Initial training, fine-tuning, and inferencing are distinct phases in the deployment of ML models for GenAI and traditional AI use cases. These phases can be correlated with appropriate venues for their execution, depending on the requirements for control and compliance over the data.

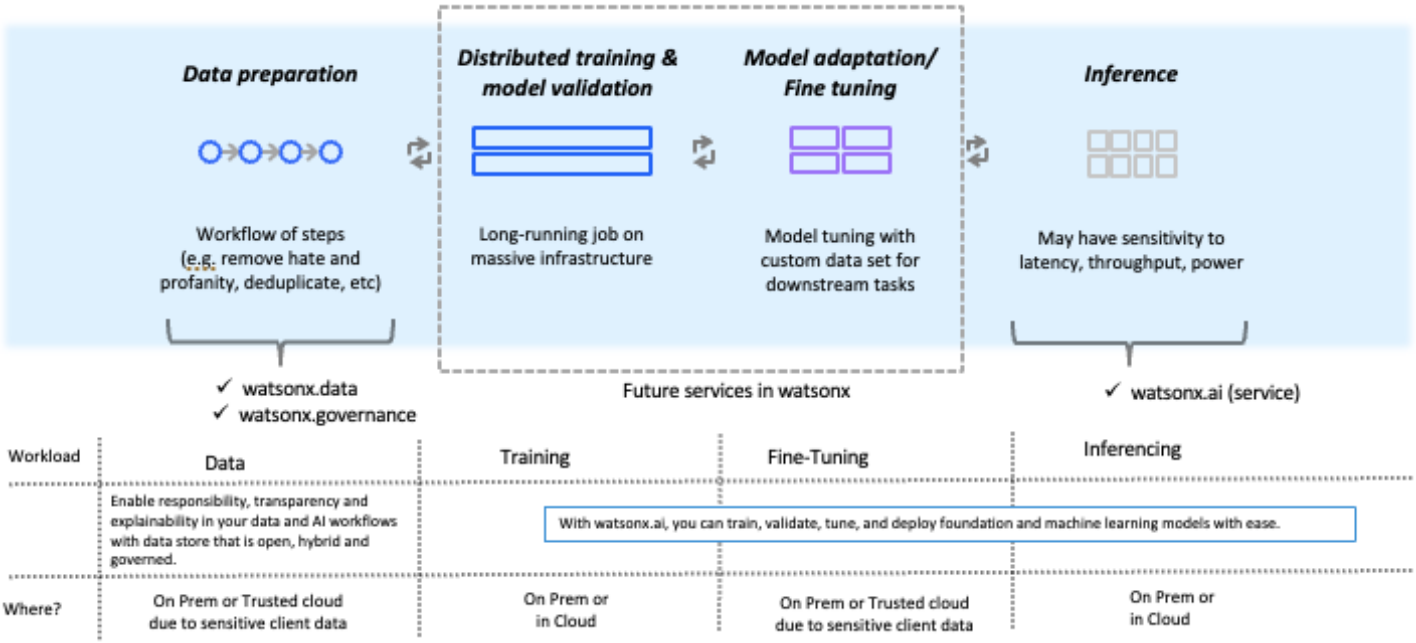

Figure 5 - watsonx Use Cases for On-premises and Trusted Cloud Environments

Leverage full-stack architecture based on VMware Cloud Foundation and Red Hat OpenShift

This full-stack architecture, built on VMware Cloud Foundation, runs Red Hat OpenShift and pairs the capabilities of the IBM watsonx platform for Gen AI and classical AI/ML-based workloads and enterprise-grade security.

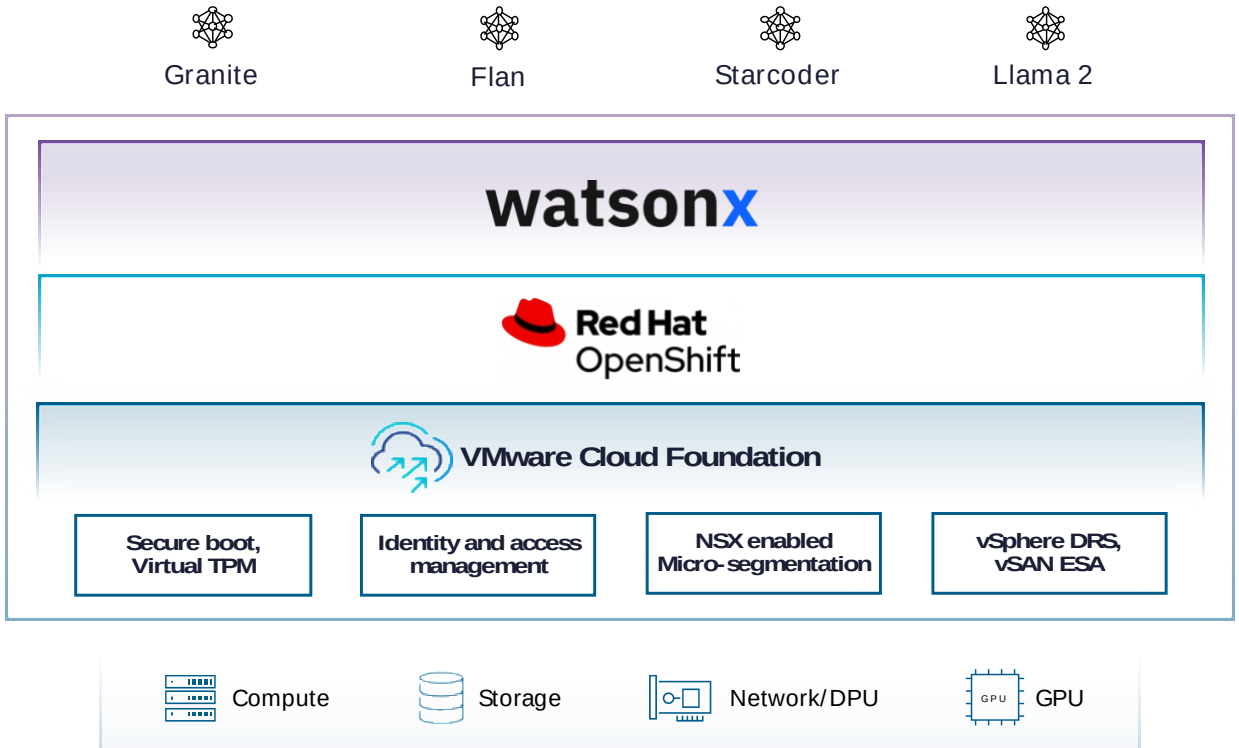

Figure 6 - VMware Private AI with IBM High-level Ecosystem

With the architecture as referenced in Figure 4, enterprises can use watsonx.ai for accessing IBM-selected open-source models from Hugging Face, as well as other third-party models and a family of IBM-trained foundation models such as Granite to support GenAI use cases and for classical AI/ML model training, validation, tuning, and deployment.

Here are a few examples of the supported models available with watsonx.ai:

- IBM foundation models: Granite (13 billion parameters)
- Open-source models: Llama 2 (70 billion parameters), Flan-t5-xxl (11 billion parameters)
- 3rd party models: StarCoder (15.5 billion parameters)

Enable secure LLM customization and fine-tuning for industry use cases

VMware Private AI with IBM can enable several use cases for enterprises by securely enabling large language models' customization, fine-tuning, and deployment (inference) within their private corporate environment. In the domain of code generation, the focus is on accelerating developer productivity while addressing critical concerns surrounding privacy and intellectual property. Moreover, in collaboration with IBM, VMware Private AI presents a significant opportunity to enhance contact center interactions. This partnership promises improved content and feedback quality for customers, resulting in more accurate responses and an overall enhanced customer experience. This partnership can significantly streamline IT operations by automating tasks like incident management, reporting, ticketing, and monitoring, ultimately saving time and effort for IT operations agents. Finally, advanced information retrieval capabilities brought about by this collaboration can elevate employee productivity by streamlining document search and policy research, fostering a more productive work environment.

watsonx.ai Installation

After completing hardware, software, storage, private container registry access and jump Linux VM with the required CLI (Cloud Pak for Data and OpenShift), here are the steps to deploy IBM WatsonX-AI on VMware: https://www.ibm.com/docs/en/cloud-paks/cp-data/4.8.x?topic=watsonxai-installing. Also, refer to the IBM watsonx.ai on VMware Cloud Foundation Reference Architecture (coming soon) for the installation process on VCF.

watsonx.data Installation

Scale analytics and AI workloads for all your data, anywhere with watsonx. data, the industry's only data store that is open, hybrid, and governed. The detailed steps can be found at: https://www.ibm.com/docs/en/SSDZ38_1.1.x/wxd/install/setup-client.html.

watsonx.governance Installation

Accelerate responsibility, and transparency and explainability in your data and AI workflows with watsonx. governance. This solution helps you direct, manage, and monitor your organization's AI activities. The detailed steps can be found at: https://www.ibm.com/docs/en/cloud-paks/cp-data/4.8.x?topic=services-watsonxgovernance.

Solution Validation

Refer to the IBM watsonx.ai on VMware Cloud Foundation Reference Architecture (coming soon) for details on the solution validation.

Conclusion

VMware Cloud Foundation delivers flexible, consistent, secure infrastructure and operations across private and public clouds. It is ideally suited to meet the demands of modern applications running on the Red Hat OpenShift Container Platform in a virtualized environment.

With VMware Cloud Foundation, we can easily manage the lifecycle of the hybrid cloud environment with a unified management plane for all applications including OpenShift. With VMware Cloud Foundation, we can leverage the leading virtualization technologies including vSphere, NSX, and vSAN.

In this solution paper, we demonstrated the architecture of running Red Hat OpenShift Container Platform with VMware Cloud Foundation. We showed the configuration details, the hardware resources, and the software resources used in the solution validation. We showed the various configuration options in addition to the best practices. VMware Cloud Foundation Manager provided lifecycle management. vSAN provided reliable, high-performance, and flexible storage to OpenShift. NSX Data Center provided the fine-grained, secured, and high-performance virtual networking infrastructure to OpenShift. Also, vSphere DRS and vSphere HA provided efficient resource usage and vSphere HA. All the above led to a consolidated solution of running the Red Hat OpenShift Container Platform with VMware Cloud Foundation.

References

- VMware Cloud Foundation
- VMware Cloud Foundation 5.1 Release Notes
- VMware vSphere
- VMware vSAN
- VMware NSX Data Center
- Red Hat OpenShift Container Platform
- IBM watsonx AI and Data Platform
- Solution YAML Files on GitHub

About the Authors

- Bryan Buckland, CTO VMware Solutions, IBM Cloud.
- Girish Manmadkar, Technology Product Manager, VMware by Broadcom
- Troy Foster, Technology Product Manager, VMware by Broadcom

Executive sponsors:

- Rajeev Bhardwaj, Broadcom
- Zane Adam, IBM
- Dilpreet Bindra, Broadcom
- Madhup Gulati, Broadcom
- Manish Aggarwal , IBM
- Shridhar Yedunuthula, Broadcom
- Remus Lazar , IBM watsonx
- Ana Manrique, IBM watsonx

Supported by:

- Fabiano Teixeira, Software Engineer, VMware by Broadcom
- Troy Lewis, IBM
- Chen Wei, Broadcom
- Yasser Sheriff, IBM
- Angad Singh, IBM

Item No: vmw-bc-vmware-validated-solution-vcf-ibm-watsonx-ai-july-2024 7/24