# Data Modernization with VMware Data Services Manager

- Enterprise-ready Data Services on VMware Cloud Foundation
- Seamless vector Database Integration for VMware Private AI Foundation with NVIDIA
- Fast and Streamlined Consumption of NVIDIA Retrieval-Augmented Generation (RAG) Workflow

vmware

## Table of contents

# Executive Summary

## Business Case

Enterprises that choose to differentiate themselves through digital transformation have an increasing appetite for speed in insight and innovation. This requires faster development and release cycles for data-driven applications. However as much of the data still stands in the legacy world, most organizations have fleets of operational databases anchored to the usual ways of provisioning and management.

By fully leveraging data to power both modern and traditional applications including Generative AI (GenAI), *VMware Data Services Manager™* helps customers accelerate their digital transformation by building, deploying, and operating a diverse and growing data estate. It offers a managed database as a service solution that brings ease of operations, automation, and scalability in *VMware Cloud Foundation™* environments. With VMware Data Services Manager, both enterprises and cloud service providers can run rich data services such as enterprise traditional databases and vector databases that are required in GenAI use cases, with other data services such as object store, streaming, caching and more on the roadmap. The deep integration with the rest of VMware Cloud Foundation components allows IT admins to use the same tools and processes that they are familiar with and deliver the right performance, reliability, and scalability in those environments.

VMware Cloud Foundation, deployed as on-premises software-defined infrastructure, serves as a basis for self-managed and provider-managed deployments by delivering full-stack private cloud platform that establishes a robust and extensible cloud operating model. It is adaptable to organizations at all stages of their multi-cloud journey that enables application mobility across geographically dispersed cloud boundaries. VMware Cloud Foundation integrates the latest virtual GPU (vGPU) technologies as part of integration with the NVIDIA AI Enterprise Suite to empower Artificial Intelligence and Machine Learning (AI/ML) programs driven by the major strategic business initiatives and modernization projects across many industry verticals.

*VMware Private AI Foundation™ with NVIDIA* deploys Retrieval-Augmented Generation (RAG) workloads with pgvector databases supported by VMware Data Services Manager to further interact and improve the quality of Large Language Models (LLMs) outputs. It also enables rapid deployment of enterprise deep learning projects with scalability and flexibility in VMware Cloud Foundation environments. The tightened collaboration between VMware and NVIDIA allows customers to scale compute and storage quickly in tandem with the customized GenAI workload needs, improve management and operational efficiency with quick action/troubleshooting, tailor to different cost and capacity request of GenAI workloads and stay in control with predictable growth, and provide monitoring and dashboards for GPU-powered applications. VMware Private AI Foundation with NVIDIA ensures privacy, security, and compliance for enterprise AI models, simplifies deployment and optimizes costs, accelerates performance and of LLMs with various choices.

In this reference architecture, we provide design and deployment guidance, performance validation, best practices for enterprise infrastructure administrators and GenAI application owners to deploy, run, and manage VMware Data Services Manager on VMware Cloud Foundation platform powered by VMware Private AI Foundation with NVIDIA.

## Business Values

The top five benefits of deploying, running, and managing VMware Data Services Manager on VMware Cloud Foundation environments powered by VMware Private AI Foundation with NVIDIA include:

- **Faster provisioning for rich data services**: Data services templates that are pre-packaged and certified by VMware and partners enable organizations to quickly spin up desired data platform.
- **Automated operations:** Common routine tasks like lifecycle management, resource handling, scaling, data protection and availability can all be automated with VMware Data Services Manager.
- **Extensive monitoring:** Gain visibility into both database and underlying local and cloud resources, including metrics such as the utilization of CPU, memory, network, local, and cloud storage, as well as curated alarms and notifications for improved visibility into the health of both the database and the underlying infrastructure.
- **Ready business for GenAI:** Enable enterprises to customize models and run GenAI applications with VMware Private AI Foundation with NVIDIA and vector database capabilities within VMware Data Services Manager.
- **Seamless integration with the cloud platform:** Empower IT organizations with enterprise-hardened data services that are deeply integrated with VMware Cloud Foundation.

## Key Results

This reference architecture is a showcase of running and managing data-as-a-service solution by VMware Data Services Manager on VMware Cloud Foundation environments. Key results can be summarized as following:

- VMware Data Services Manager delivers enterprise hardened modern data services with developer-friendly consumption, automated management, and unified control and visibility.
- The enterprise-ready MySQL, PostgreSQL, and Google Omni AlloyDB (tech-preview) data services provided with scalable and consistent performance, easy-to-deploy high availability, as well as on-demand business continuity and data recovery.
- VMware Data Services Manager provides pgvector as vector database capability to accelerate enterprise adoption of AI/ML workloads within an optimized virtualization infrastructure on VMware Cloud Foundation environment.
- VMware Private AI Foundation with NVIDIA enables GenAI applications with better content generation and diversity of responses by seemless integration with pgvector databases supported by VMware Data Services Manager.

## Audience

This solution is intended for IT administrators, infrastructure experts, data scientists and GenAI specialists who are involved in the early phases of planning, design, and deployment of modern data services on VMware Cloud Foundation. It is assumed that the reader is familiar with the concepts and operations of VMware Data Services Manager, VMware Cloud Foundation, NVIDIA RAG workflows and related components.

# Technology Overview

Solution technology components are listed below:

- VMware Cloud Foundation
- VMware Data Services Manager
- VMware Private AI Foundation with NVIDIA

## VMware Cloud Foundation

VMware Cloud Foundation provides a ubiquitous cloud platform for both traditional enterprise and modern applications. Based on a proven and comprehensive software-defined stack including VMware vSphere®, VMware vSAN™, VMware NSX®, VMware vSphere with VMware Tanzu®, and VMware Aria Suite™ and VMware Data Services Manager, VMware Cloud Foundation provides a complete set of software-defined services for compute, storage, network, container, and cloud management. The result is agile, reliable, efficient cloud infrastructure that offers consistent operations across private and public clouds. VMware Cloud Foundation also offers a unified platform for managing all workloads, including VMs, containers, and AI technologies, through a self-service and automated IT environment.

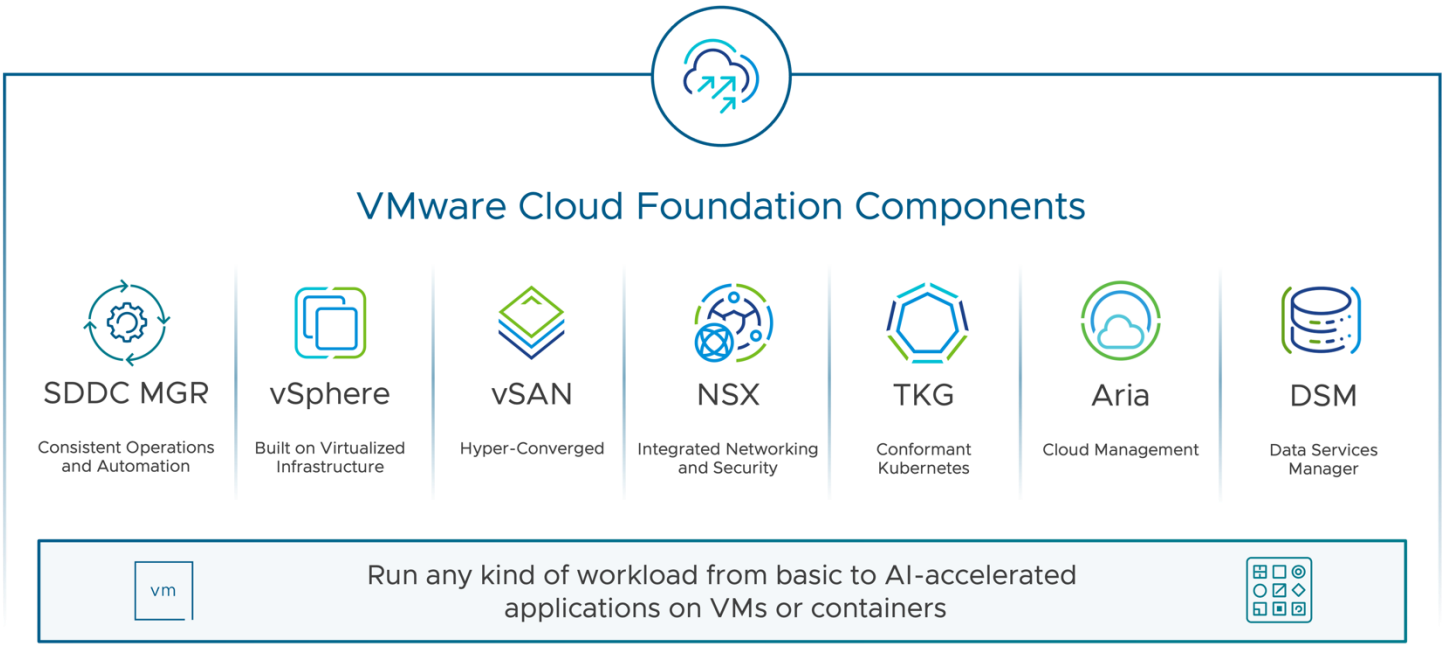


Figure 1. VMware Cloud Foundation Components

## VMware Data Services Manager

Now part of VMware Cloud Foundation, VMware Data Services Manager empowers IT organizations with key VMware Cloud Foundation on native data services that are enterprise-hardened and easy to manage at scale. VMware Data Services Manager also supports pgvector extension to improve the quality of enterprise LLM outputs with RAG workflows. With the deployment of vector databases using pgvector on PostgreSQL, customers get better context and diversity of responses for their GenAI applications.

VMware Data Services Manager helps our customers to:

- Increase Control for VI Admin through a purpose-built solution for VMware Cloud Foundation with high availability, performance, backup, and monitoring.
- Reduce cost and complexity for Data Team through pre-built automation to deploy and manage data services at scale.
- Accelerate velocity for Developers through self-service, rich APIs, and a range of certified DB offerings.

## VMware Private AI Foundation with NVIDIA

VMware collaborates with NVIDIA to offer VMware Private AI Foundation with NVIDIA that enables enterprises to leverage LLMs, produce more secure and private models for their internal usage, enable enterprises to offer GenAI as a service to their users, and more securely run inference workloads at scale. This integrated GenAI platform enables enterprises to run RAG workflows, fine-tune and customize LLM models, and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns. It simplifies GenAI deployments for enterprises by offering an intuitive automation tool, deep learning VM images, vector database, and GPU monitoring capabilities.

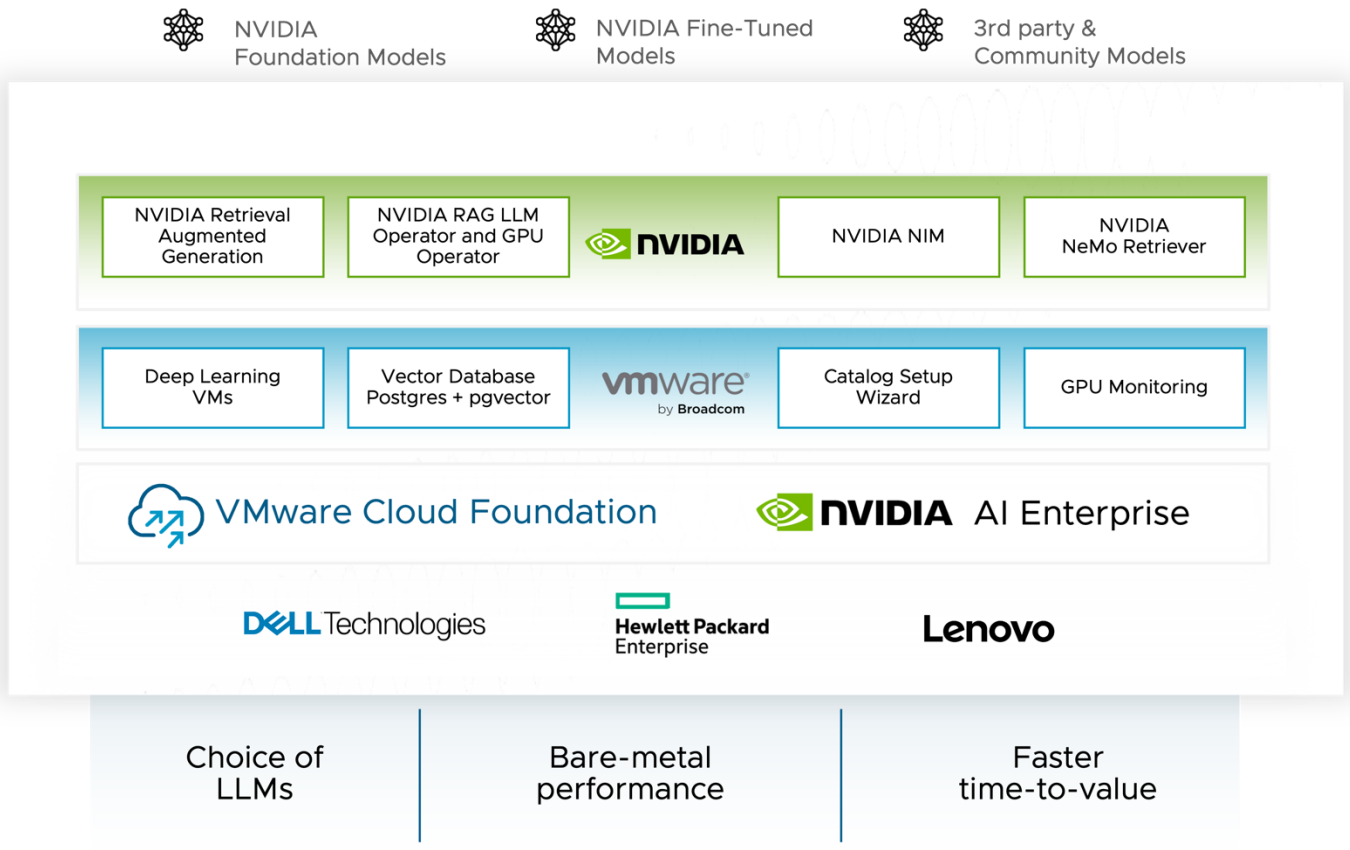


Figure 2. Introducing VMware Private AI Foundation with NVIDIA

## Solution Configuration

This section introduces the following resources and configurations:

- Architecture diagram
- Hardware resources
- Software resources
- Virtual machine configuration
- vSAN configuration
- Deploying and configuring VMware Data Services Manager
- Configuring the Object Storage
- Provisioning Data Services

## Architecture Diagram

In the solution, we deployed different data engines provided by VMware Data Services Manager on VMware Cloud Foundation platform, including MySQL, PostgreSQL, and Google AlloyDB Omni (currently in Tech Preview). VMware Data Services Manager requires S3-compatible object storage for provider repository, logs, and backups.

VMware Data Services Manager provides self-managed and optimized data services for traditional and modern application that requires MySQL or PostgreSQL database. VMware Data Services Manager also supports pgvector extension within PostgreSQL database which was fully certified under VMware Private AI Foundation with NVIDIA. As shown in Figure 3, we validated NVIDIA RAG pipelines with pgvector integration provided by VMware Data Services Manager.

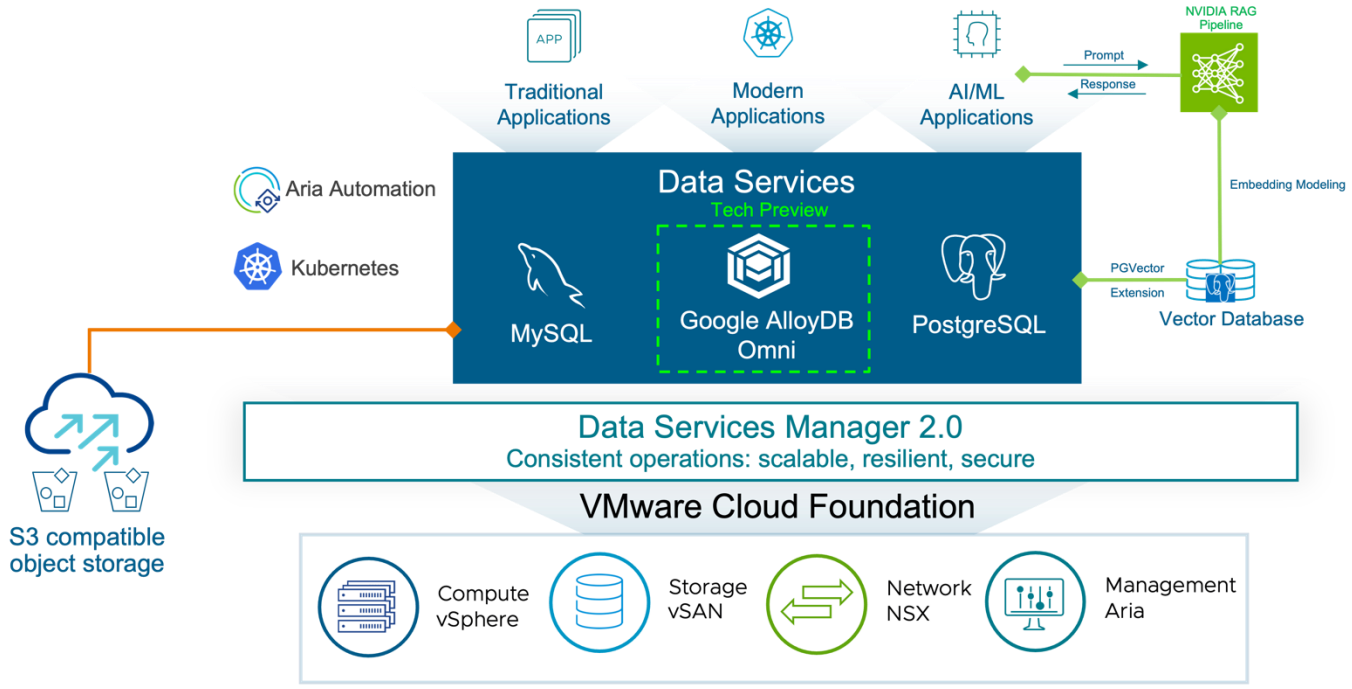


Figure 3. Architectural Diagram

## Hardware Resources

**Table 1. Hardware Configuration**

| PROPERTY | SPECIFICATION |
|---|---|
| Server model name | 6 x Dell PowerEdge R640 |
| CPU | Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz, 28 cores each |
| GPU | NVIDIA A100 Tensor Core GPU |
| RAM | 512GB |
| Network adapter | Mellanox Technologies ConnectX-4 Lx EN NIC 25GbE dual-port SFP28 |
| Disks | 6 x Dell Express Flash NVMe P4610 1.6TB PCIe SSD each host |

## Software Resources

Table 2 shows the software resources used in this solution.

**Table 2. Software Resources**

| SOFTWARE | VERSION | PURPOSE |
|---|---|---|
| VMware Cloud Foundation | 5.1.1 | VMware Cloud Foundation is the leading cloud platform for both traditional enterprise and modern applications. |
| VMware Data Services Manager | 2.0.1 | Modern database and data services management for vSphere that provides a data-as-a-service toolkit for on-demand provisioning and automated management of PostgreSQL and MySQL databases. |
| VMware Private AI Foundation with NVIDIA | / | It is a component of VMware Cloud Foundation. VMware Private AI Foundation with NVIDIA enables enterprises to run RAG workflows, fine-tune and customize LLM models, and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns. |

## Virtual Machine Configuration

The virtual machine configuration in this solution is described in Table 3.

**Table 3. Virtual Machine Configuration**

| VM Role | vCPU | Memory (GB) | VM Count |
|---|---|---|---|
| Provider | 8 | 16 | 1 |
| MySQL | 32 | 128 | 1 |
| PostgreSQL | 32 | 128 | 1 / 3 / 5 |
| Database Client | 4 | 16 | 1 |
| MinIO Linux | 4 | 16 | 1 |

## vSAN Configuration

vSAN provides default storage policy with erasure coding enabled. In this solution, we used “vSAN Express Storage

Architecture (ESA) Default Policy – RAID 6" as the storage policy for data services virtual machines and object storage virtual machine deployed on vSAN ESA. The testbed is configured with 6 physical hosts and the ESA RAID 6 is the optimum storage policy that provides better resiliency with FTT=2 for database workloads, equivalent performance with no compromise as compared to RAID 1, and better space-efficiency (1.5x). Deploying the object storage on vSAN can also benefit from the flexible vSAN storage policy tailored by different SLA requirements.

## Deploying and Configuring VMware Data Services Manager

VMware Data Services Manager simplifies the deployment and configuration process with tighter integration with vSphere and vSAN. The following enhancement can be found in the current release:

- VMware Data Services Manager now supports single OVA deployment for the provider virtual machine without agent requirement.
- VMware Data Services Manager is registered as a vSphere plugin to ease the installation procedure.
- Introduction of Infrastructure Policies to allow vSphere administrator to configure, manage, and monitor the database infrastructure within VMware vCenter® console.
- Introduction of Kubernetes in VMware Data Services Manager to support cloud-native modern services by leveraging Cluster API Provider vSphere (CAPV) and different data service operators.
- Self-service layer for developers to provision data services with Aria Automation and Kubernetes tools.
- New declarative API available to manage the infrastructure and databases within VMware Data Services Manager.

Figure 4 shows the infrastructure policies created for the test environment with VMware Data Services Manager plugin installed in the vCenter console.

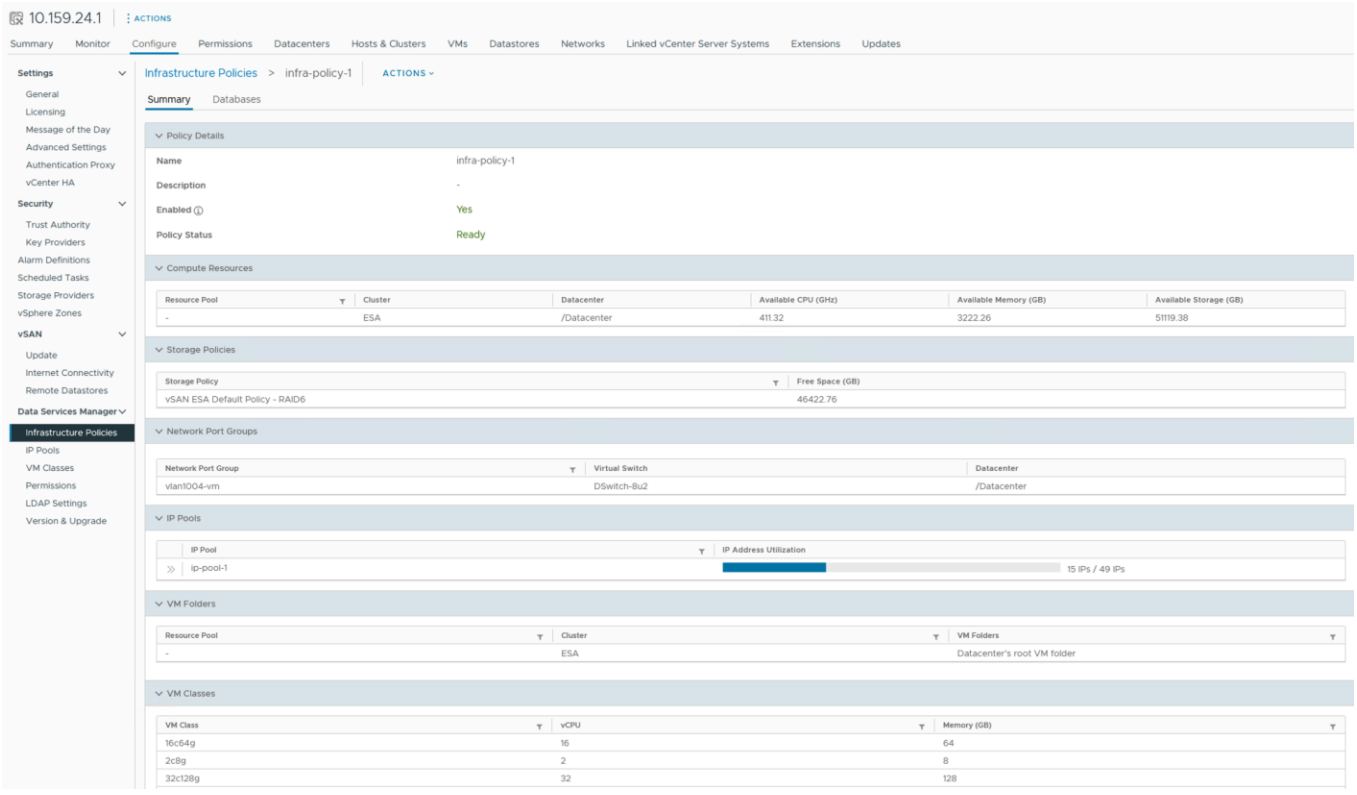


Figure 4. Infrastructure Policies introduced in VMware Data Services Manager

VMware Data Services Manager provides an interactive web console through which the users can provision, manage, and monitor the data services. Refer to the *VMware Data Services Manager documentation* for more details.

## Configuring the Object Storage

VMware Data Services Manager uses S3-compatible object storage to maintain local copies of database templates, software updates, log bundles, and database backups. The following storage repositories are required to configure properly.

- Provider Repo: storage for database templates and software updates.
- Provider Log Repo: storage for log bundles of provider virtual machine.
- Provider Backup Repo: storage for backup of the internal vPostgreSQL database of the provider.
- Database Backup Repo: storage for data service backups.

It is required to configure TLS enabled endpoints for each of the storage locations required by VMware Data Services Manager. In this solution, we deployed MinIO Object Storage for Linux in a Ubuntu 22.04 virtual machine and created multiple S3 buckets for the data repository as shown in Figure 5. The single node MinIO object storage deployment may also take advantage of vSAN to ensure data reliability and availability at storage layer for provider repositories, database templates, and backups in VMware Data Services Manager.

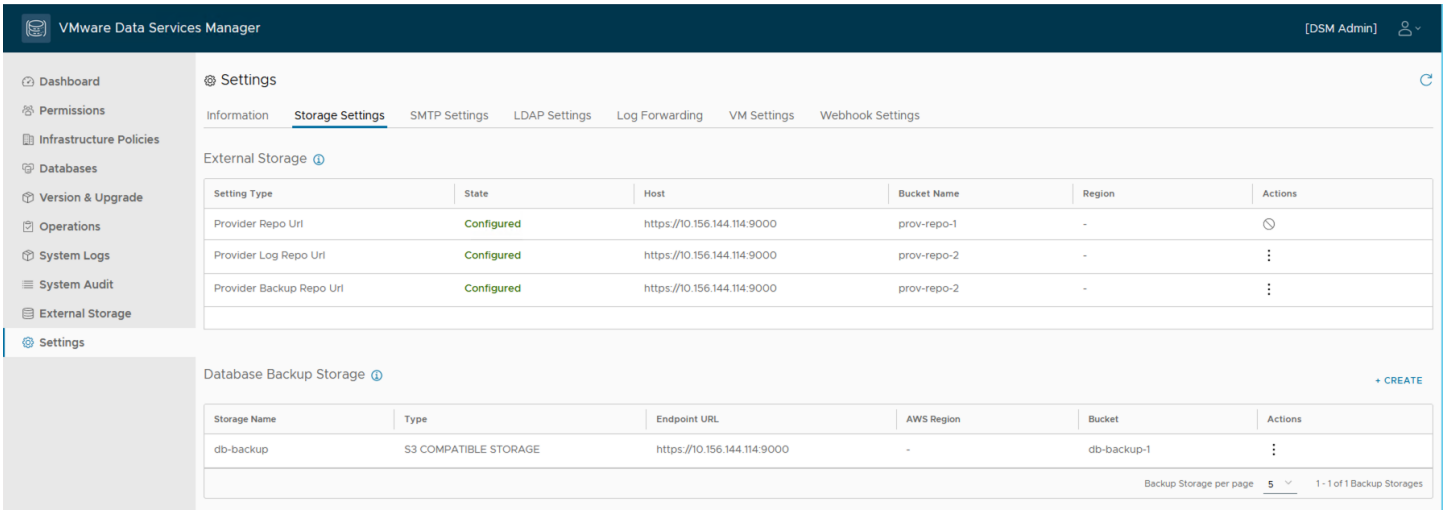


Figure 5. Configuring the Object Storage in VMware Data Services Manager

## Provisioning Data Services

You may provision MySQL and PostgreSQL databases using the pre-configured templates as provided by VMware Data Services Manager. It releases certified VMware Data Services Manager database templates and software updates to *Tanzu Network* and the administrator must *configure the Tanzu Network Refresh Token* properly. Otherwise, you may refer to *Manually Populating Database Templates and Software Updates* for air-gapped environment to set up the data services. After the data services are enabled in the console, you can proceed to deploy the database using the desired infrastructure policy and backup/maintenance strategy.

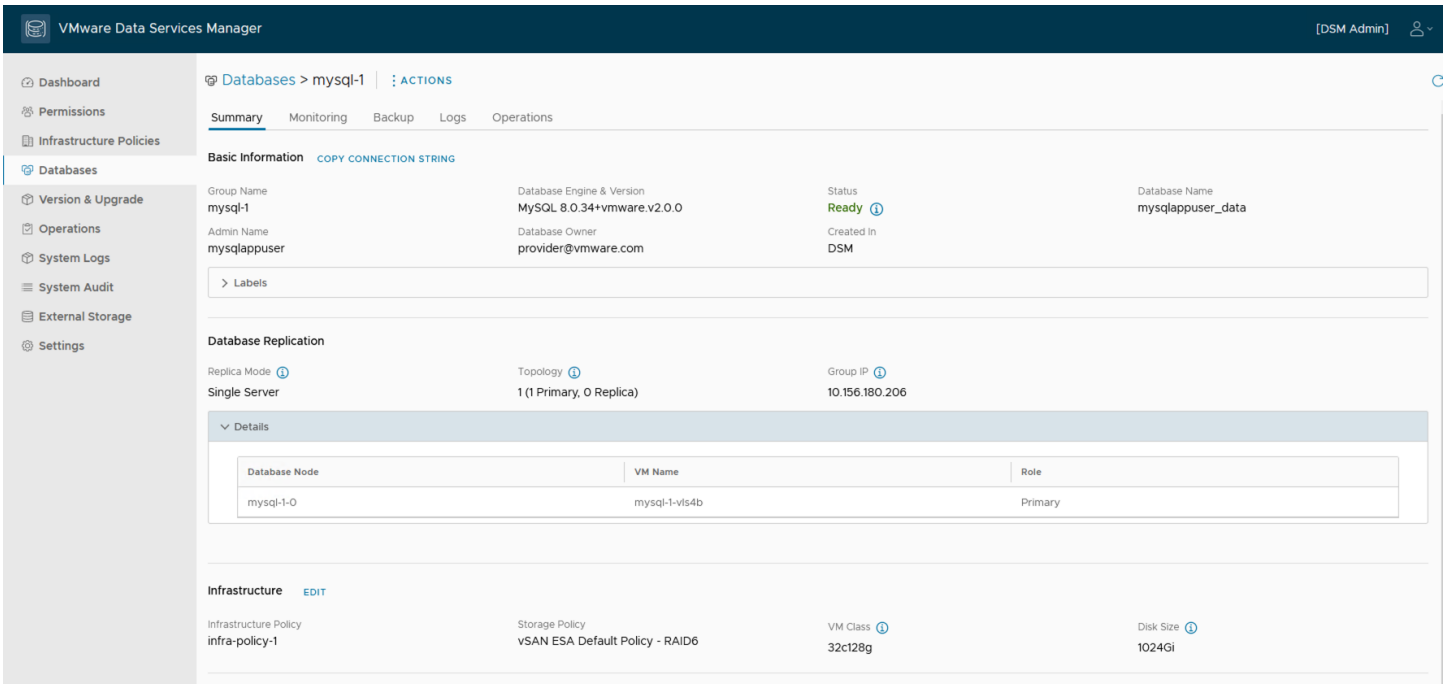


Figure 6. Provisioning MySQL in VMware Data Services Manager

VMware Data Services Manager also provides tech preview for Google AlloyDB Omni data service. This feature is not enabled by default and customers should contact Google and VMware by Broadcom for AlloyDB Omni support in VMware Data Services Manager. Refer to *Data Service Manager Documentation* for more details.

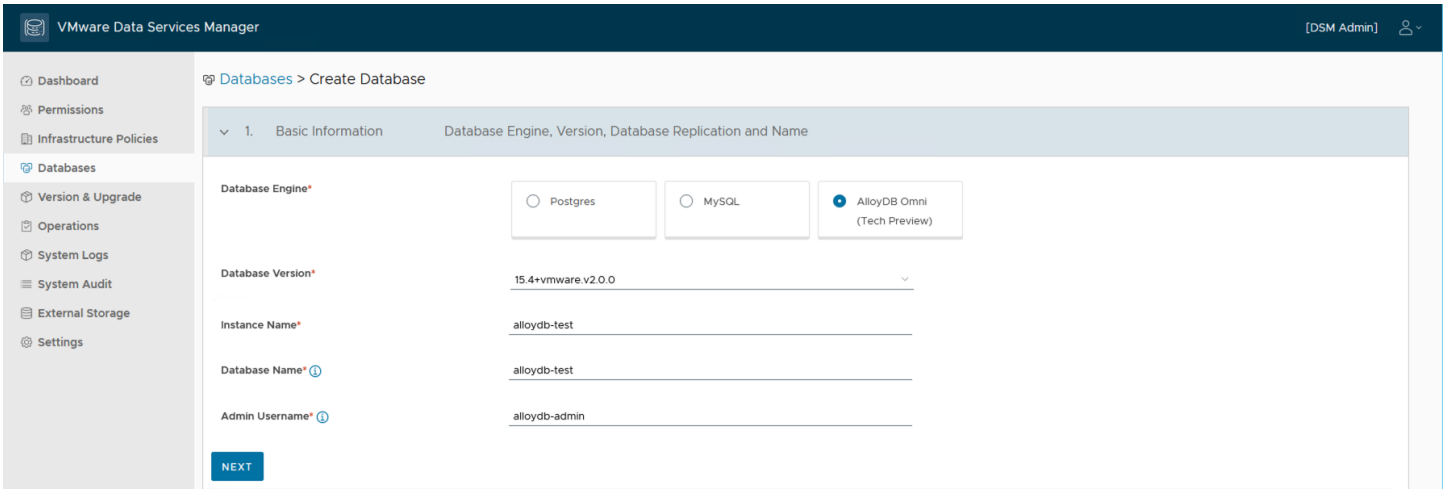


Figure 7. Google AlloyDB Omni as Technical Preview in VMware Data Services Manager

## Solution Validation

### Test Overview

We deployed MySQL and PostgreSQL database using VMware Data Services Manager and validated multiple use cases by different benchmark tools. The following test cases are included in this solution:

- MySQL performance and scalability test: We validated a MySQL database provisioned by VMware Data Services Manager and tested the performance capability and scalability.
- PostgreSQL high availability test: VMware Data Services Manager supports high availability configuration for PostgreSQL database and we validated the compromise of performance and data availability.
- Vector database test: We validated the large-dimensional vector database performance of pgvector extension provided by VMware Data Services Manager using VectorDBBench.
- Sample use case for VMware Data Services Manager integration with VMware Private AI Foundation with NVIDIA.

## Test Tools

We used the following monitoring tools and benchmark tools in the solution testing:

- Monitoring tools

  **vSAN Performance Service**

  *vSAN Performance Service* is used to monitor the performance of the vSAN environment, using the vSphere web client. The performance service collects and analyzes performance statistics and displays the data in a graphical format. You can use the performance charts to manage your workload and determine the root cause of problems.

  **vSAN Health Check**

  *vSAN Health Check* delivers a simplified troubleshooting and monitoring experience of all things related to vSAN. Through the vSphere web client, it offers multiple health checks specifically for vSAN including cluster, hardware compatibility, data, limits, physical disks. It is used to check the vSAN health before the mixed-workload environment deployment.

  **VMware Data Services Manager Web Console**

  VMware Data Services Manager collects health and metric data for each database which can be viewed under the web console to track resource consumption, performance, and activity of the databases. Figure 8 shows the database metrics under monitoring tab of the testing database.

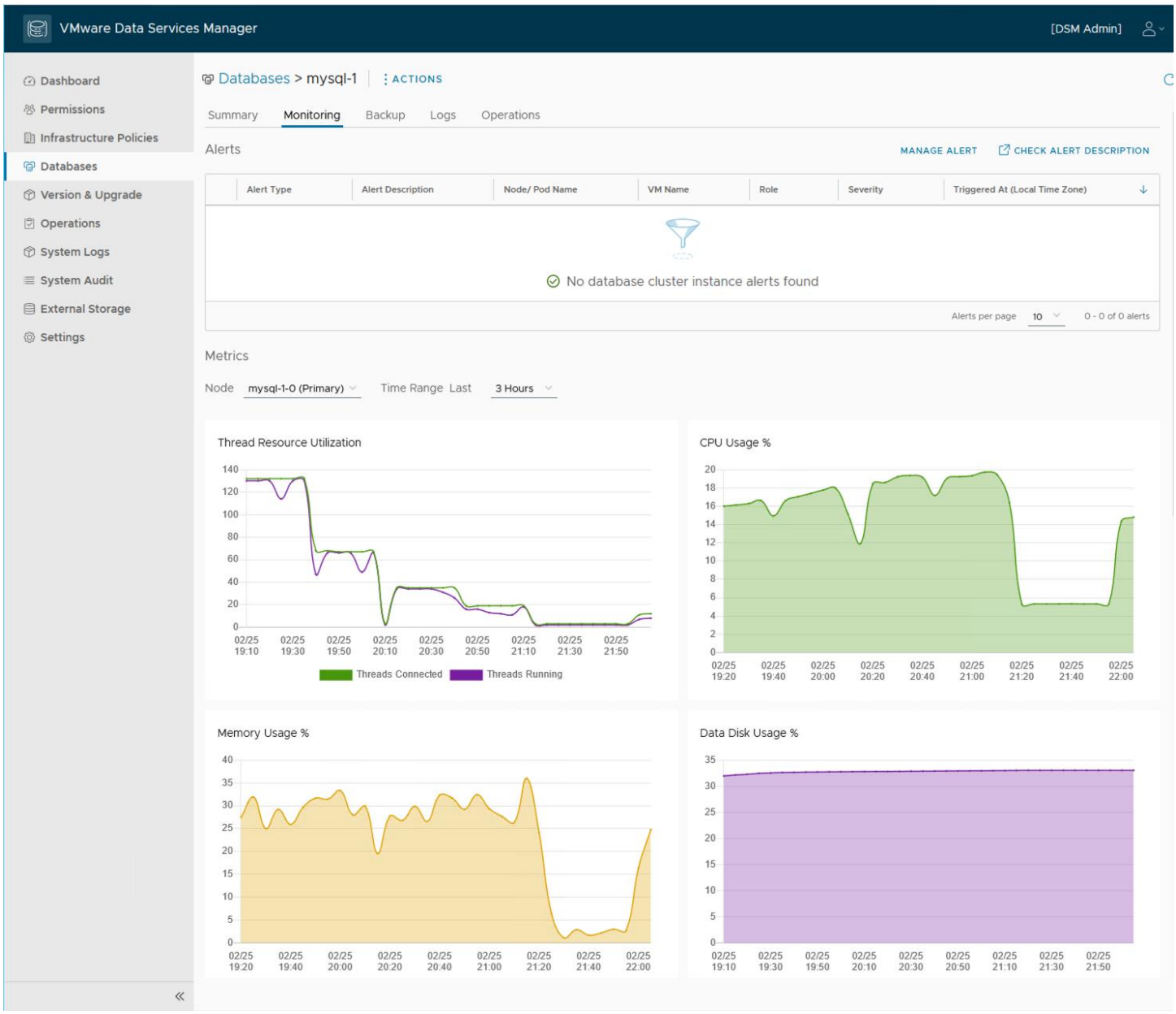


Figure 8. Monitoring Console in VMware Data Services Manager

- Database benchmark tool

***SysBench***

SysBench is a modular, cross platform, and multithreaded benchmark tool for evaluating OS parameters that are important for a system running a database under intensive load. The Online Transaction Processing (OLTP) test mode is used in solution validation to benchmark a real MySQL database performance.

In this solution validation, we used the SysBench OLTP library to populate MySQL and PostgreSQL test databases and generate workload for performance and scalability test.

***VectorDBBench***

VectorDBBench is a go-to tool for the ultimate performance and cost-effectiveness comparison for mainstream vector databases. It helps set up diverse testing scenarios including insertion, searching, and filtered searching by closely mimicking real-world production environments.

In this solution validation, we loaded 1M dataset from actual production scenarios such as Cohere and OpenAI into pgvector extension and tested the search performance and filter search performance cases to demonstrate the vector database performance capability.

### MySQL Database Configuration

The MySQL VM Class was configured with 32 virtual CPU (vCPU) and 128GB Memory with a 1TB data disk for database. VMware Data Services Manager provides a set of balanced parameter settings for MySQL database; for example, the default innodb_buffer_pool_size is configured at a default level of about 50 percent (58GB out of 128GB).

VMware Data Services Manager also supports pre-deployment or post-deployment parameter tunings for the databases in the Advanced Settings available in the web console. We provide a set of parameters designed for performance profile in the performance and scalability testing. Table 4 summarizes the database options configured for MySQL.

Table 4. MySQL Database Profile

| MySQL Default Profile | MySQL Performance Profile |
|---|---|
| Innodb_bufffer_pool_size=58GB | innodb_buffer_pool_size=96G<br>innodb_flush_log_at_trx_commit=2<br>innodb_io_capacity=5000<br>innodb_io_capacity_max=50000<br>innodb_log_buffer_size=32M<br>innodb_log_file_size=24G<br>innodb_read_io_threads=64<br>innodb_write_io_threads=64 |

**Note:** The performance profile was applied to the performance and scalability test only. The parameter values should be subject to adjust based on the actual VM class and database use cases in production environment. You may refer to *MySQL InnoDB Startup Options and System Variables documentation* for more details.

### MySQL Performance and Scalability Test Result

**Test findings:** Linear performance scalability of MySQL database provided by VMware Data Services Manager.

We populated 10 tables with each containing 100 million records using SysBench benchmark tool as MySQL test database. Table 5 lists the test database size.

Table 5. Database Size of MySQL OLTP Performance Scalability Test

| Database Platform | Table Count | Rows per table | Actual Size (including index) |
|---|---|---|---|
| MySQL | 10 | 100,000,000 | 204GB |

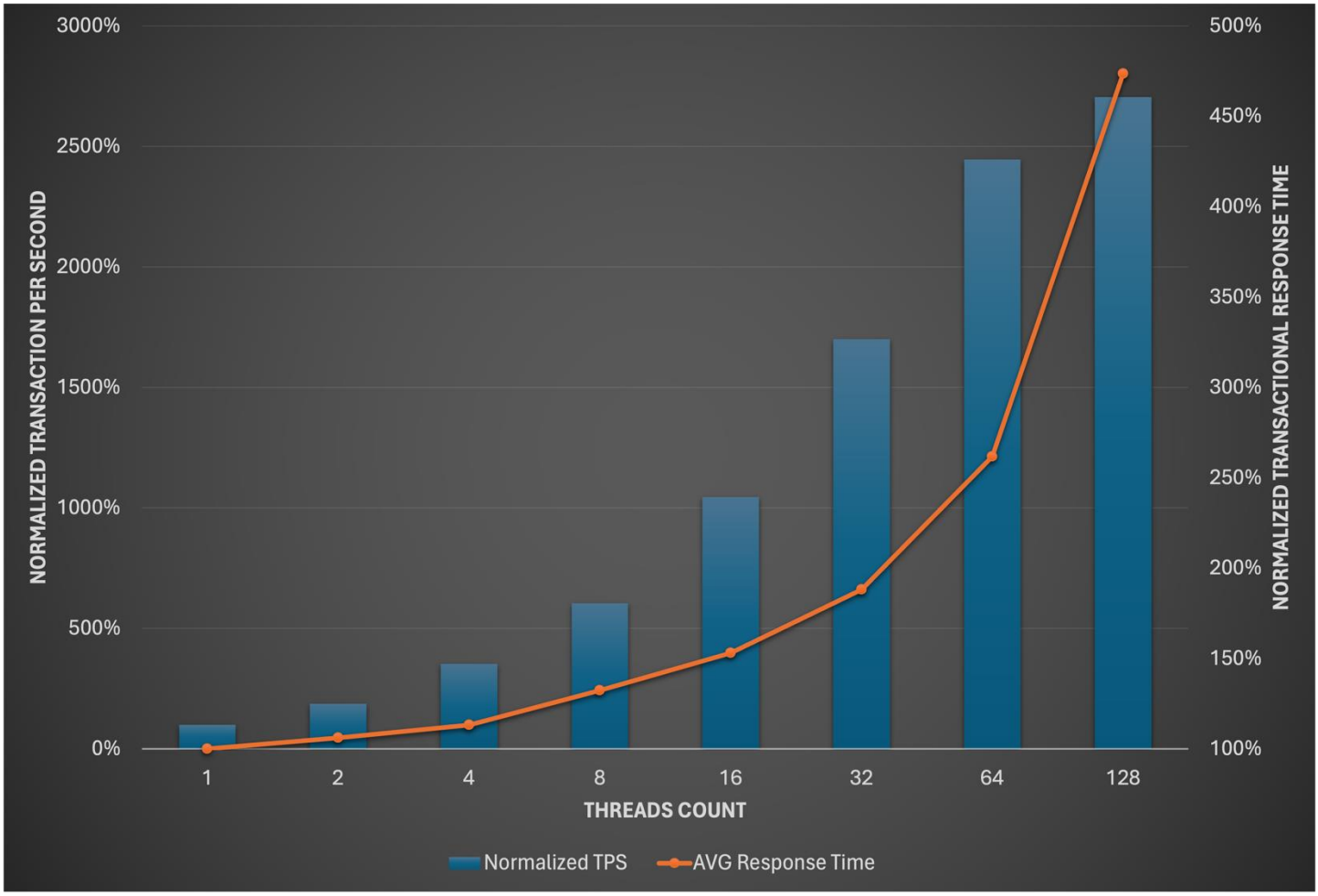


Figure 9. MySQL OLTP Performance and Scalability Test Result

Figure 9 shows the MySQL scalability test result with the performance profile. The result is measured by normalized transaction per second and transactional response time. The performance profile helps optimize the MySQL buffer pool and log file with better multi-threading capability for both read and write requests. With 128 SysBench threads, it achieved 27x higher Transaction per Second (TPS) compared to a single thread with average response time increased by 4.7x.

Table 6 shows the test results for both default profile and performance profile for 128 threads count. With the MySQL performance profile, it achieved 5.18x better normalized TPS and reduced the average transactional response time by over 80 percent. The parameter optimization of performance profile can be easily achieved with the database options in the advanced configuration for the data service.

Table 6. MySQL Performance Scalability Test Results: Default vs Performance Profile

| Metrics | Default Profile | Performance Profile |
|---|---|---|
| Normalized TPS | 1x | 5.18x |
| Average Transactional Response Time | 100% | 19.29% |

## PostgreSQL Database Configuration

The PostgreSQL VM Class was configured with 32 vCPU and 128GB Memory with a 1TB data disk for database. VMware Data Services Manager sets the default shared_buffers of PostgreSQL to 25 percent of the virtual machine memory size and configured the

max_connections to 100. The user may further customize the database options in the advanced configuration in VMware Data Services Manager. We tested the following scenarios for PostgreSQL:

- Single server mode with standalone PostgreSQL. The optimized parameters are applied in the testing.

  Table 7. PostgreSQL Database Profile

| PostgreSQL Default Profile | PostgreSQL Optimized Profile |
|---|---|
| shared_buffers=20518MB<br>max_connections=100 | effective_cache_size=2097152<br>max_connections=200<br>max_wal_size=24576<br>shared_buffers=6710886<br>wal_buffers=262143 |

- Single vSphere cluster with one primary, one replica, and one monitor node for PostgreSQL cluster.
- Single vSphere cluster with one primary, three replicas, and one monitor node for PostgreSQL cluster.

### PostgreSQL High Availability Test Result

Test findings: PostgreSQL cluster with synchronous replication in VMware Data Services Manager provides higher data availability at the database level while maintaining a minimal performance compromise.

We populated 10 tables with each containing 100 million records using SysBench benchmark tool as test database for PostgreSQL database. Table 8 lists the test database size.

Table 8. Database Size of PostgreSQL OLTP High Availability Test

| Database Platform | Table Count | Rows per Table | Actual Size (including index) |
|---|---|---|---|
| PostgreSQL | 10 | 100,000,000 | 238GB |

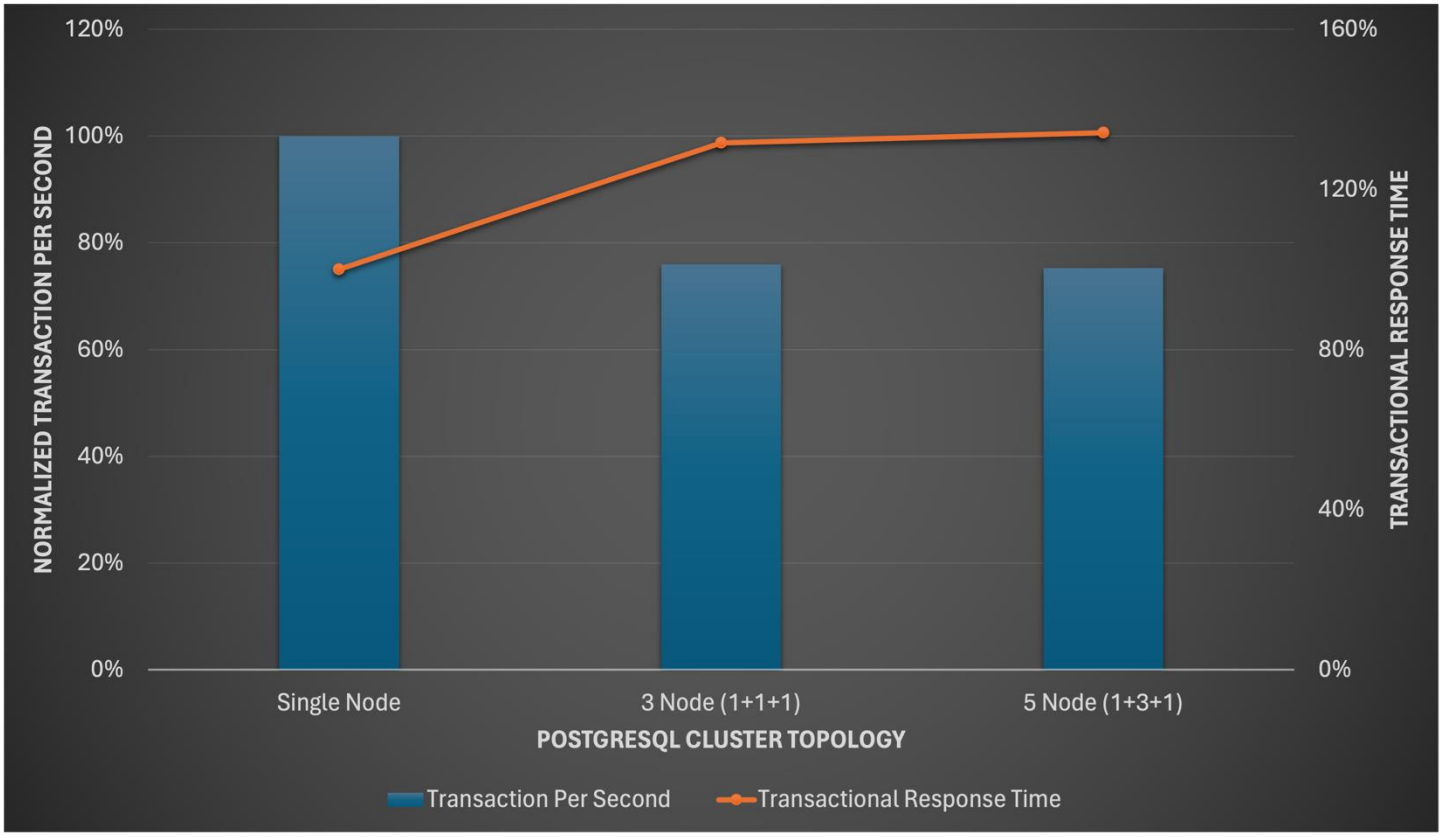


Figure 10. PostgreSQL High Availability Test Results

PostgreSQL cluster in VMware Data Services Manager supports synchronous replication from the primary to the replica node. The monitor is responsible for orchestrating the failover. A static kube-vip load balancer service is deployed to redirect the read-write requests from the client side to the primary node. To ensure best performance, it is recommended to place the kube-vip on the same node as the primary node to get the optimized network path.

Figure 10 shows the benchmark result of OLTP read/write workload with 128 threads for PostgreSQL cluster of different topologies. We drove up to 70 percent VM CPU utilization in the single node testing and collected the result as baseline. We then scaled the single node PostgreSQL cluster into a three-node topology with one primary, one replica, and one monitor node. We ran the same OLTP benchmark using SysBench and measured the performance result. The three-node cluster result was about 75% of the single node in terms of TPS while maintaining the high availability on the database level to tolerate one failure. The transaction response time was up by about 30 percent. By further scaling out to a 5-node topology, we did not monitor further performance degradation whereas it provided up to 2 concurrent node failures.

In summary, the high availability architecture of the PostgreSQL cluster in VMware Data Services Manager has some performance compromise because of synchronous replication enabled among the nodes but consistent as expected. Knowing that trade-off between performance and data protection is inevitable, the higher level of data availability maintains the business continuity for critical enterprise data services workloads.

## Vector Database Benchmark Test Result

**Test findings:** The pgvector extension in VMware Data Services Manager provides the vector database required to store high-dimensional vector embeddings generated by machine learning models and algorithms. With the tailored pgvector virtual machine sizing, the performance demonstrated highly scalable, consistent, and predictable in terms of QPS, recall, and more metrics.

VMware Data Services Manager provides pgvector extension to support vector database capability. VectorDBBench is the benchmark tool for mainstream vector databases and a go-to tool for the performance and cost-effectiveness comparison. The following use case were tested in this solution.

- Search Performance Test (Cohere 1M vectors, 768 dimensions) - Performance768D1M
- Search Performance Test (OpenAI 500K vectors, 1536 dimensions) - Performance1536D500K

The test metrics are measured as follows:

- QPS: Query per second, the higher the better
- Recall: Used to calculate the percentage of relevant results returned by a query, the higher the better.

Benchmark parameters tuning as follows:

- Lists: A good starting point of lists is number of rows divided by 1,000 for up to 1M rows and sqrt(rows) for over 1M rows. We used 1,000 in the test.
- Probes: A good starting point of probes is square of the number of lists. In the test, we used 10 for better result of QPS and 40 for better result of recall.

Figure 11 and Figure 12 measured the QPS and recall with different VM configuration with linear performance capability in terms of Cohere dataset and OpenAI dataset, respectively. The pgvector VM performance with VectorDBBench in terms of query per second scaled linearly as we increased the number of CPU and memory.

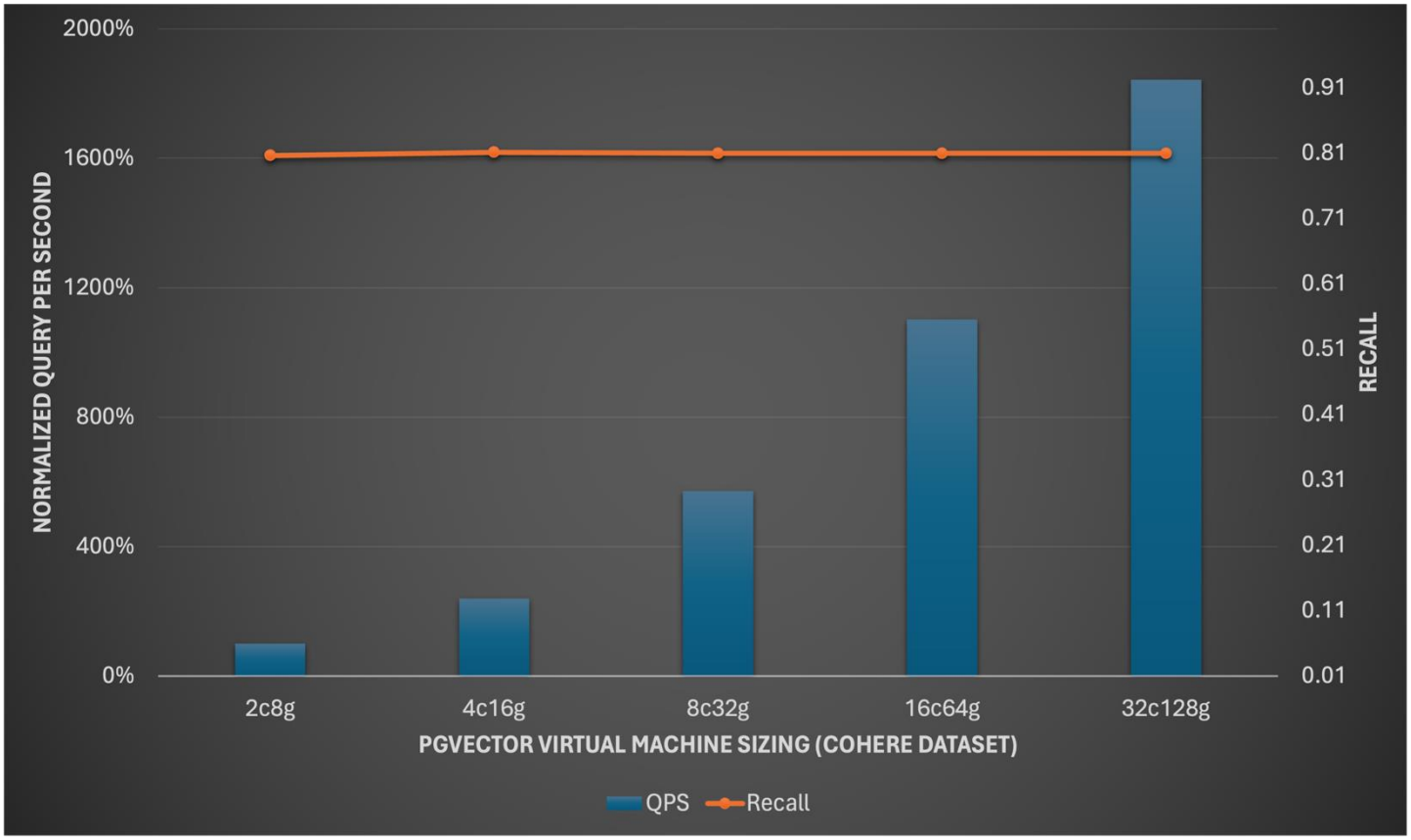


Figure 11. QPS and Recall for Cohere Dataset: VM T-shirt Sizing

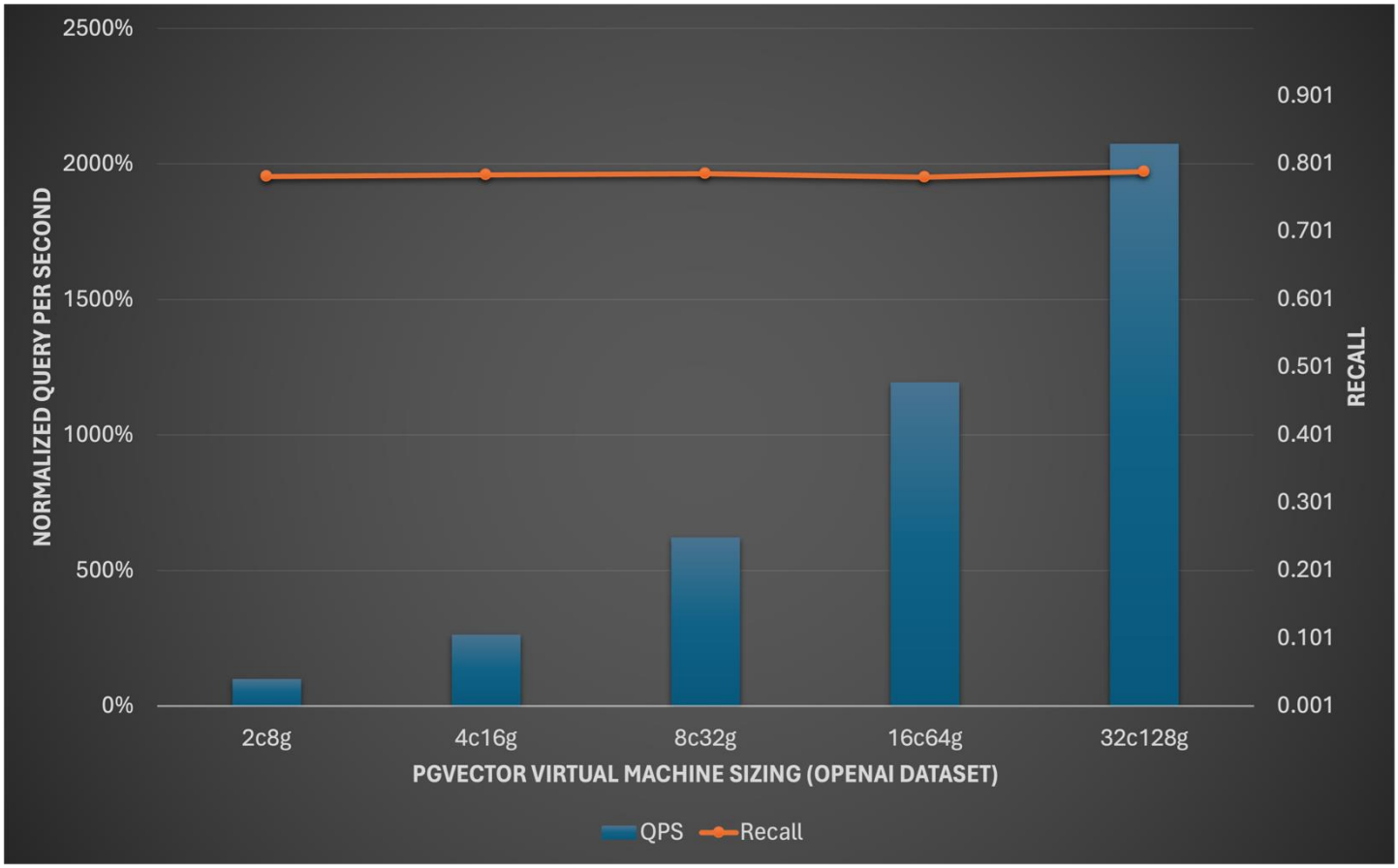


Figure 12. QPS and Recall for OpenAI Dataset: VM T-shirt Sizing

Table 9 shows the sample test result for better QPS result of the embeddings with lists=1,000 and probes=10. The PostgreSQL virtual machine installed with pgvector extension was configured with 32 vCPU and 128GB memory. We pushed the CPU utilization of the database virtual machine to over 90 percent with 35 concurrent benchmark user loads.

Table 9. Test Results: Lists=1000 and Probes=10 (For better QPS)

| Benchmark Settings | Test Case | QPS | Recall |
|---|---|---|---|
| Lists=1,000<br>Probes=10 | Performance768D1M | 417.52 | 0.7888 |
| | Performance1536D500K | 583.6291 | 0.8124 |

Table 10 shows the test result for better recall value with lists=1000 and probes=40. A higher value of recall means more accuracy as per query, but the trade-off is lower QPS result. It is recommended to tune both lists and probes for your workloads based on the type of training data.

Table 10. Test Results: Lists=1000 and Probes=40 (For better recall)

| Benchmark Settings | Test Case | QPS | Recall |
|---|---|---|---|
| Lists=1000<br>Probes=40 | Performance768D1M | 118.3484 | 0.9317 |
| | Performance1536D500K | 180.8529 | 0.9356 |

## Use Case Study—VMware Data Services Manager with NVIDIA RAG

As we demonstrated above with the VMware Data Services Manager capability to accommodate the modern GenAI workloads, it enables privacy, security, and compliance for those AI models with an integrated GenAI platform on VMware Cloud Foundation. Figure 13 shows a sample workflow of NVIDIA RAG pipeline for typical enterprise applications, the vector database can be deployed with pgvector in VMware Data Services Manager. For more details, refer to *Deploying a Vector database in VMware Private AI Foundation with NVIDIA*.

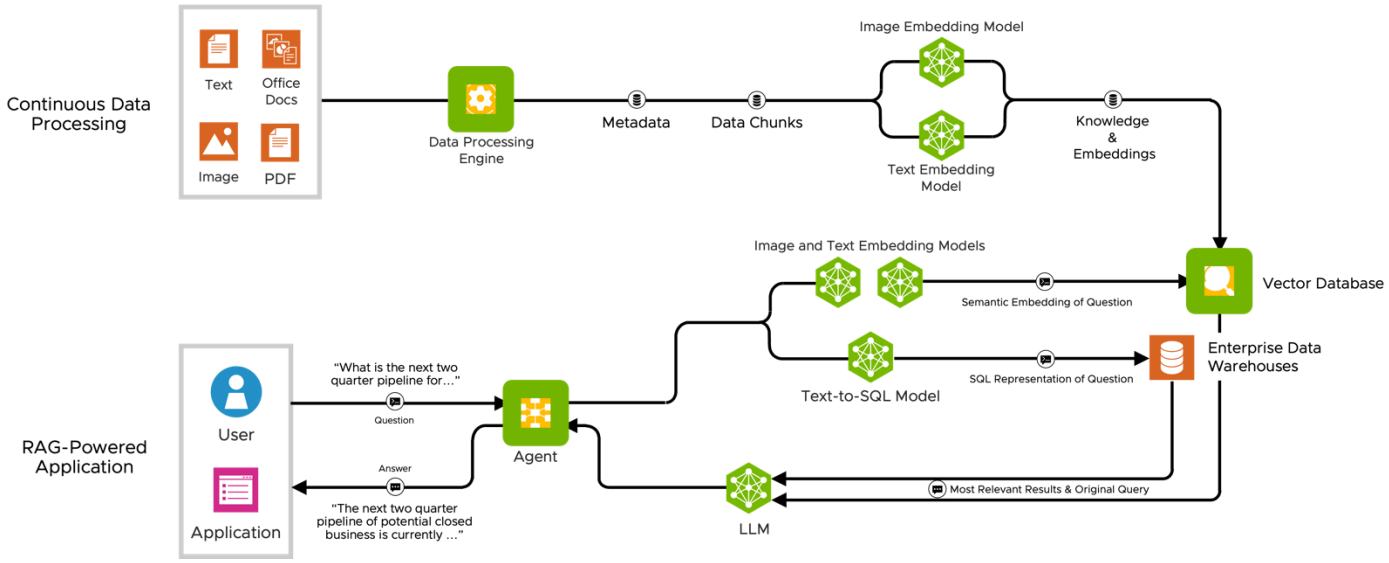


Figure 13. RAG Pipeline Example in VMware Private AI Foundation with NVIDIA

### Sample Use Case 1: Developer RAG on Deep Learning Virtual Machine

VMware Private AI Foundation with NVIDIA provides enterprise-ready deep learning virtual machine with optimized and enhanced GPU configuration. The deep learning virtual machine is a quick starting point for users to adopt the RAG process in the VMware Cloud Foundation environments.

Figure 14 shows a sample LLM playground deployed by the NVIDIA developer RAG examples on GPU enabled deep learning virtual machine powered by pgvector in VMware Data Services Manager. For more details, refer to *Deploy a Deep Learning VM with a RAG Workload*.

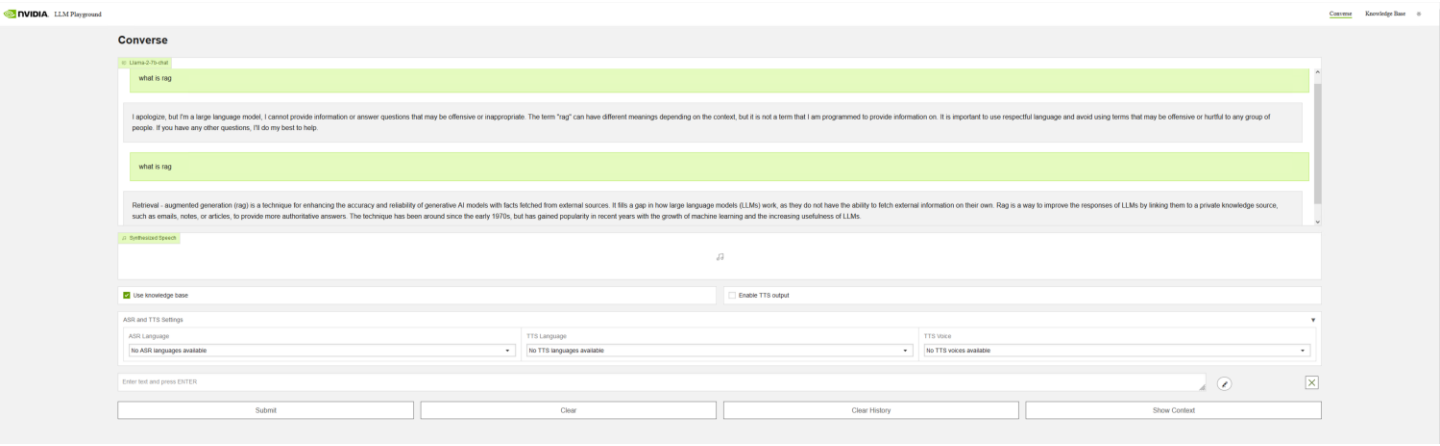


Figure 14. Sample Chatbox Powered by NVIDIA RAG Examples and VMware Data Services Manager

### Sample Use Case 2: Enterprise RAG on VMware Tanzu Kubernetes Grid Services

To further empower the GenAI workloads on VMware Private AI Foundation with NVIDIA, the enterprise RAG pipeline allows users to deploy customizable RAG pipelines through VMware Tanzu Kubernetes Grid™ Service in VMware Cloud Foundation environments. It aims to meet the more comprehensive demand of large enterprise customers on GenAI solutions to improve productivity, efficiency, and profit in diversified use cases. For more details, refer to *Deploy a RAG Workload on a TKG Cluster*.

**Note:** The NVIDIA RAG examples in the sample use cases can be found *here*.

## Best Practices

In this solution, we validated the performance scalability and high availability of the data platforms in VMware Data Services Manager and the pgvector support for the vector database adoption in the deployment of GenAI workloads.

The following recommendations provide the best practices and sizing guidance for VMware Data Services Manager on VMware Cloud Foundation environments.

- Enterprise data services performance:
  - For MySQL best practice, refer to *MySQL on VMware Data Services Manager*
  - For MySQL parameter tuning, refer to *InnoDB Startup Options and System Variables*
  - For PostgreSQL best practice, refer to *PostgreSQL on VMware Data Services Manager*
- Postgres high availability: VMware Data Services Manager supports 3-node and 5-node topology for PostgreSQL cluster. The leader node that hosted the default kube_vip load balancer service is a random activity within the Kubernetes world. For best performance consideration, it is recommended to place the kube_vip leader node on the same node as the PostgreSQL cluster primary node. Currently there is no best way to force the kube_vip placement. The following code lists an alternative to move the leader node, which may require to run multiple times to ensure the kube_vip land on the PostgreSQL primary node.

  On the provider VM:

  # ssh -i /opt/vmware/tdm-provider/provisioner/sshkey capv@<ip-of-node-running-kube-vip-leader>

  On the kube_vip leader node:

  $ sudo su –

  # mv /etc/kubernetes/manifests/kube-vip.yaml .

  // Wait 20 seconds

```
# mv kube-vip.yaml /etc/kubernetes/manifests/
```

Verify the leader node place on the same node as PostgreSQL cluster primary node:

```
# KUBECONFIG=workload_cluster.yaml kubectl get lease -n kube-system plndr-svcs-lock -o yaml
apiVersion: coordination.k8s.io/v1
kind: Lease
metadata:
  creationTimestamp: "2024-03-08T03:30:41Z"
  name: plndr-svcs-lock
  namespace: kube-system
  resourceVersion: "878446"
  uid: b309b340-9c84-438e-a8bc-83d7df856c18
spec:
  acquireTime: "2024-03-11T06:28:08.547493Z"
  holderIdentity: pgtest2-dnp8s     //holderIdentity field indicates the current leader node of the PostgreSQL cluster
  leaseDurationSeconds: 15
  leaseTransitions: 6
  renewTime: "2024-03-11T06:28:08.554799Z"
```

- Infrastructure Policies:
  - IP Pools: Prepare and assign enough static IP addresses for VMware Data Services Manager.
  - VM Class: A good starting point for CPU/Memory ratio is 1:4 for general data services workload use case.
  - User Permission: Plan VMware Data Services Manager Admin and VMware Data Services Manager User accordingly for local users and LDAP users.
- Object Storage:
  - Use TLS enabled endpoint for object storage URL required in VMware Data Services Manager.
  - Carefully prepare the provider repository which does not allow to change post-configuration.
  - For air-gapped environment, refer to *Manually Populating Database Templates and Software Updates* to set up the object storage and database templates.
- Pgvector:
  - To enable pgvector extension, visit https://github.com/pgvector/pgvector?tab=readme-ov-file#getting-started.
  - VMware Data Services Manager helps maintain the pgvector version management. The current support pgvector version is 0.4.4 as of 2.0 release. VMware is actively working on updating pgvector to the latest release with more vector database capability support.
  - IVFFLAT is currently the index type that supported by pgvector in VMware Data Services Manager. It divides the index into lists and searches a subset of those lists that are closest to the query vector. For best practices of lists and probes setting for IVFFLAT index, refer to *https://github.com/pgvector/pgvector?tab=readme-ov-file#ivfflat.*
- GenAI:
  - GPU related GenAI workload reference architecture on VMware platform:*Deploying Enterprise-Ready Generative AI on*

*VMware Private AI*

- *VMware Private AI Foundation for NVIDIA Official Documentation*

- vSAN consideration:
  - Use Erasure coding RAID-5/6 as the default storage policy for vSAN ESA as it eliminates the trade-off of performance and deterministic space efficiency. Select FTT=1 using RAID-5 and FTT=2 using RAID-6 depending on the number of hosts present in the ESA cluster and your data availability requirement.
  - ESA can support up to 32 maximum snapshots in a chain. As a best practice, always maintain a short snapshot chain for management and storage consideration. Refer to *Best practices for using VMware snapshots in the vSphere environment* for more details.

## Conclusion

VMware Data Services Manager provides modern data services with self-service and fleet management capabilities that are tightly integrated with vSphere and VMware Cloud Foundation environments. It supports one-click deployment as extension, rich, and simplified data services operations and lifecycle management. VMware Data Services Manager also offers easy-to-use consumption operator via the customized Kubernetes platform and API ready for deep customization as the application team requires.

In this solution, we demonstrated the performance scalability and data high availability for MySQL and PostgreSQL data services that are provisioned by VMware Data Services Manager. The result showcases a linear scalability with predictable and consistent performance for those enterprise-class database workloads while maintaining high data availability with minimal performance impact.

The pgvector extension in VMware Data Services Manager enables vector database integration that are widely used in modern GenAI use cases. This solution demonstrated pgvector enriched results by running the common benchmark tool to evaluate vector database with the real-world training dataset. The integration with VMware Private AI Foundation with NVIDIA for pgvector further allows enterprise customers to quickly adopt the best-fit RAG pipeline to leverage the LLMs for different use case requirements such as code generation, contact centers resolution, IT operations automation, and advanced information retrieval.

## Reference

- *VMware Cloud Foundation*
- *VMware Data Services Manager*
- *Announcing Initial Availability of VMware Private AI Foundation with NVIDIA*
- *VMware Cloud Foundation AI/ML Solutions*
- *NVIDIA RAG Documentation*

## About the Author

Mark Xu, Product Marketing Engineer of VMware Cloud Foundation in VMware by Broadcom, wrote the original version of this paper.

The following reviewers also contributed to the paper contents:

- Christos Karamanolis, Software Engineer of VMware Cloud Foundation in VMware by Broadcom
- Michael Gandy, Product Marketing Engineer in VMware by Broadcom
- Qi Liu, Software Engineer of VMware Cloud Foundation in VMware by Broadcom
- Frank Che, Software Engineer of VMware Cloud Foundation in VMware by Broadcom
- Michael West, Product Marketing Engineer in VMware by Broadcom
- David Manconi, Sales in VMware by Broadcom
- Thomas Sauerer, Staff Cloud Solution Architect in VMware by Broadcom
- Kim Delgado, Solution Architect in VMware by Broadcom
- Robert Eckhardt, Technology Product Manager in VMware by Broadcom
- Yu Wang, Technology Product Manager in VMware by Broadcom
- Ting Yin, Product Marketing Engineer in VMware by Broadcom

- Chen Wei, Senior Manager of the Workload Technical Marketing team in VMware by Broadcom
- Catherine Xu, Manager of the Workload Technical Marketing team in VMware by Broadcom


Item No: vmw-bc-wp-temp-a4-word-