Implementing a Digital Assistant with Red Hat OpenShift AI on Dell APEX Cloud Platform for Red Hat OpenShift

H19818

November 2023

White Paper

Abstract

This white paper provides a summary for using Red Hat OpenShift AI on Dell APEX Cloud Platform to create a digital assistant. This solution leverages a Large Language Model (LLM) and the Retrieval Augmented Generation (RAG) technique in combination with a set of vectorized documents.

Copyright

Contents

Introduction

Executive summary

Artificial Intelligence (AI), deep learning (DL), and machine learning (ML) have quickly risen to the top of the commercial and organizational priority list. These applications have emerged as a pivotal force for reshaping industries, processes, and customer experiences. AI applications are revolutionizing the way organizations leverage data. By deploying AI-powered algorithms, organizations can extract valuable insights from vast datasets, enabling data-driven decision-making. These applications can analyze data at high speeds and massive scale, identifying patterns and trends that would otherwise be impossible. In doing so, AI applications transform data into actionable intelligence, providing organizations with a competitive edge in a data-centric world.

Some examples of AI use cases include:

- Conversational agents and chatbots for customer service
- Audio and visual content creation
- Software programming
- Security, fraud detection, and threat intelligence
- Natural language interaction and translation
- Supply chain management
- Medical imaging and diagnostics

Dell APEX Cloud Platform for Red Hat OpenShift is designed collaboratively with Dell Technologies and Red Hat to optimize and extend OpenShift deployments on-premises with an integrated operational experience. By combining Dell's expertise in delivering robust infrastructure solutions with Red Hat's industry leading OpenShift Container Platform, this collaboration empowers organizations to start on a transformative journey towards modernization and innovation.

Document purpose

This white paper provides readers with an overview of how to create a digital assistant using Red Hat OpenShift AI on Dell APEX Cloud Platform for Red Hat OpenShift.

Audience

This white paper is intended for AI solutions architects, data scientists, data engineers, IT infrastructure managers, and IT personnel who are interested in, or considering, implementing AI/ML deployments.

Business and technical challenges

AI and ML are transformative technologies, but they come with business and technical challenges. Addressing the key challenges outlined in the following sections is critical for a successful AI/ML implementation.

Data quality and availability

AI and ML models rely on data, and ensuring data quality, accuracy, and availability can be challenging. Many organizations struggle with siloed data, incomplete datasets, and data that may not adequately represent real-world scenarios.

Infrastructure requirements

AI and ML applications have to scale to handle extremely large datasets and accommodate for increases in user loads. Efficiently maintaining high performance can be a challenge for organizations that do not have the proper infrastructure.

Model complexity

Developing and managing complex algorithms and models requires a robust skill set and the right staff to maintain the environment.

Ethical considerations

AI and ML models can have ethical implications as they can inadvertently introduce bias and discrimination into decision-making processes. Addressing ethical implications and ensuring fairness in algorithms is a growing challenge.

Cost management

AI and ML workloads can be resource-intensive and lead to high infrastructure and operational costs. It is essential to optimize resource usage to minimize costs while maintaining operational efficiency.

Regulatory compliance

Some industries and applications have strict regulatory requirements that they must follow. These restrictions can impact what data is included in a model and where the model runs (such as on-premises, colocation, cloud).

Solution benefits

AI and ML benefits

AI/ML offers numerous benefits across various domains and industries. Some key benefits include:

- **Automation**: AI can automate repetitive and time-consuming tasks, increasing efficiency and reducing the requirement for manual labor.
- **Data analysis**: ML can predict future outcomes based on historical data, enable organizations to proactively address issues, make forecasts, and mitigate risks.
- **Fraud detection and security**: ML models can detect fraudulent transactions and identify security threats by analyzing patterns and anomalies in data.
- **Cost savings**: Automation and predictive maintenance powered by AI can reduce operational costs, improve process efficiency, and extend the lifespan of equipment.
- **Improved customer service**: Chatbots and virtual assistants powered by AI provide 24/7 Customer Support, improving customer satisfaction.
- **Improved productivity:** AI and ML can automate routine and repetitive tasks allowing employees to focus more on high-level tasks, improving operational efficiency, and cost-effectiveness.

Dell and Red Hat partnership

The partnership between Dell and Red Hat gives customers the following advantages for deploying AI and ML workloads:

- A full-stack turnkey platform for AI/ML workloads powered by Dell infrastructure and software, with NVIDIA accelerators, and Red Hat OpenShift Data Science software on top of OpenShift Container Platform.
- Validated designs that are thoroughly tested with proven configurations to reduce customers' time and effort, accelerating time to value. These integrated solutions have been documented to help speed-up and simplify deployment of new applications.
- Accelerated digital transformations by enabling organizations to take the guesswork out of deploying AI/ML innovative solutions.
- Engineering validated AI/ML solutions to meet specific use cases, running on-premises, hybrid cloud, or as-a-service.

Partner technology overview

Red Hat

Red Hat is the industry-leading provider of enterprise open-source solutions—including Linux, cloud, container, and Kubernetes. Red Hat delivers hardened solutions that make it easier for enterprises to work across platforms and environments, from the core data center to the network edge. It enables enterprises to harness the power of open-source software in their digital transformations. Red Hat also empowers organizations to modernize their IT infrastructures, embrace hybrid and multicloud environments, and accelerate innovation while maintaining stability and security.

Solution overview

Data science is the process of extracting meaningful insights from data through a combination of advanced data analytics and machine learning techniques. In today's data-driven world, organizations need data science to thrive and remain competitive. It empowers businesses to unlock the hidden value within their vast datasets, enabling them to make informed decisions, optimize operations, enhance customer experiences, and drive innovation.

This solution consists of Red Hat OpenShift AI running on Dell APEX Cloud Platform for Red Hat OpenShift to illustrate how Large Language Models (LLMs) and the Retrieval Augmented Generation (RAG) framework can seamlessly process, store, and deploy large-scale models efficiently and cost-effectively in the form of a digital assistant.

Digital assistants are prevalent across a wide range of verticals and use cases. The technology is designed to assist users by answering questions and processing simple tasks. By anchoring the model with relevant documentation, answers remain up to date and contain information unique to the organization.

PDF files or web pages are divided into smaller chunks and embeddings. These embeddings are stored in a vector database like Redis. Results from the vector database are ranked and sent to the Llama 2 model when users submit a query a semantic search is performed against the vector database. The Llama 2 model answers the queries based on results from the vector database and its pretrained capabilities.

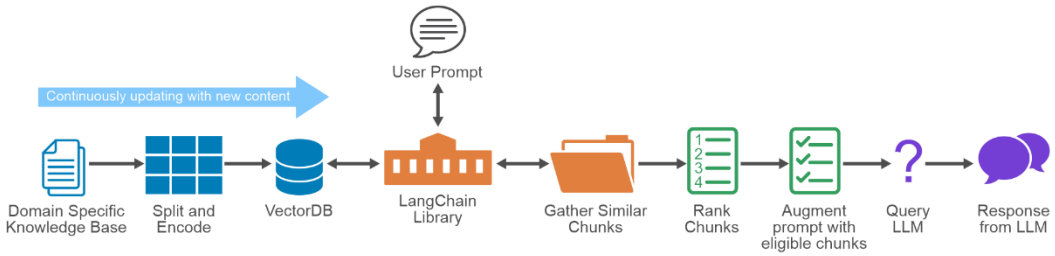

Figure 1. Digital assistant workflow

Dell APEX Cloud Platform for Red Hat OpenShift

The Dell APEX Cloud Platform for Red Hat OpenShift introduces a new level of integration for running OpenShift on bare metal servers. Until now, all infrastructure has been managed through separate OEM tools and separately from the management of OpenShift. This requires IT staff with unique skill sets to maintain the system, increasing operational costs.

The Dell APEX Cloud Platform Foundation software mitigates this complexity by integrating the infrastructure management into the OpenShift Web Console. This integration enables administrators to update the hardware using the same workflow that updates the OpenShift software. It also enables OpenShift administrators to manage the infrastructure using the same management tools they use to control the cluster and the applications that run on it.

The benefits of the Dell APEX Cloud Platform for Red Hat include:

- **OpenShift on bare metal:** Eliminates virtualization layer to reduce both acquisition and operation costs, and management overhead.
- **Full automation**: Brings cloud-like agility and efficiency for cloud-native workloads on bare metal through software-driven automation with full life cycle management.
- **Common storage across the OpenShift ecosystem:** Eliminates storage silos across OpenShift ecosystem through optimized and consistent storage, improving operational efficiency and enabling an enhanced hybrid experience.
- **New standard in OpenShift hybrid computing:** Cloud-integrated management and economics set a new standard for edge-core-cloud infrastructure, forming the new de-facto building block for OpenShift hybrid cloud ecosystem.

For more information, see the Dell APEX Cloud Platforms for Red Hat OpenShift webpage.

The figure below shows a sample of the OpenShift Web Console.

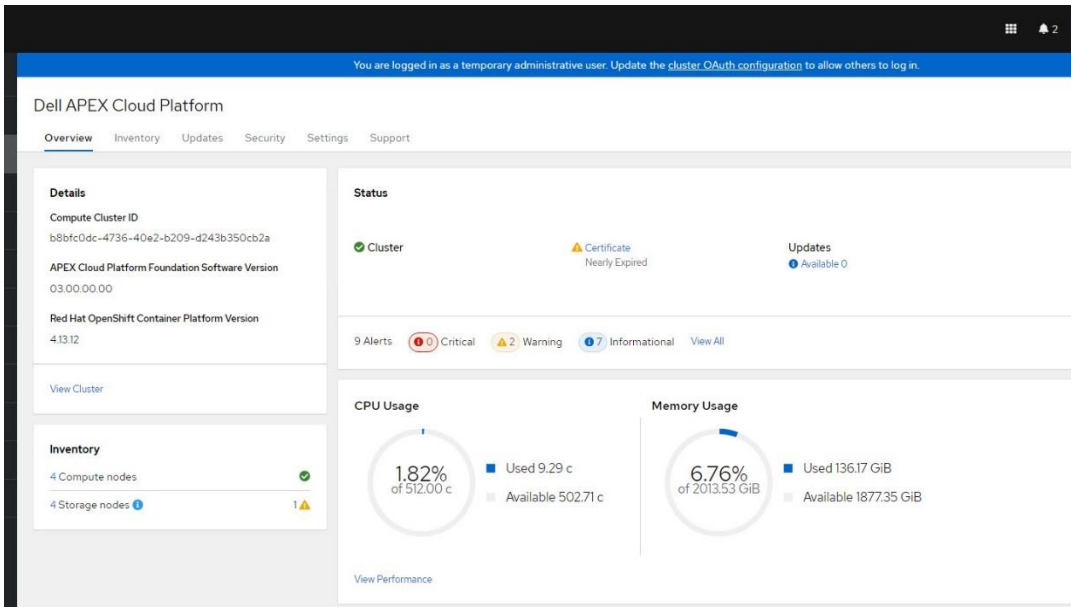

Figure 2. Red Hat OpenShift web console

Dell storage

The APEX Cloud Platform for Red Hat OpenShift uses separate storage nodes to provide persistent storage for the compute cluster. This separation enables the compute and storage nodes to scale independently. The storage cluster is based on Dell PowerFlex software-defined storage architecture. This disaggregation allows customers with existing PowerFlex storage to deploy only the compute cluster. Customers can opt for full integration; in which case the OpenShift Web Console would be used to control both the compute and storage clusters.

OpenShift Container Platform

Red Hat OpenShift Container Platform is the industry-leading hybrid cloud application platform powered by containers and Kubernetes. Using OpenShift Container Platform simplifies and accelerates the development, delivery, and life cycle management of a hybrid mix of applications across on-premises, public clouds, and edge environments.

OpenShift Container Platform is designed to deliver continuous innovation and speed at any scale, helping organizations to be ready for today and build for the future, this includes:

- Modernizing existing applications
- Developing new cloud-native applications
- Integrating data analytics and AI/ML capabilities to achieve data driven insights
- Integrating software from independent software vendors (ISVs) and cloud providers

OpenShift Data Science

Red Hat OpenShift AI is an open-source ML platform for the hybrid cloud. By providing a fully supported environment to establish MLOPs best practices, data scientists and developers can rapidly train, deploy, and monitor ML workloads and models on-premises and in the public cloud.

OpenShift AI combines Red Hat components, open-source software, and technology partner offerings with the flexibility to develop and serve models on-premises or in public clouds. OpenShift AI is available as an add-on cloud service to Red Hat OpenShift Dedicated and Red Hat OpenShift Service on AWS or as a self-managed software product.

Red Hat OpenShift AI offers organizations an efficient way to deploy an integrated set of common open-source and third-party tools to perform AI/ML modeling. The platform makes it simple to embrace hardware acceleration, including 4th Generation Intel Xeon Scalable Processors and NVIDIA GPU infrastructure, without requiring users to perform daily management of Kubernetes.

Object storage

Using object storage for LLMs and datasets is crucial in modern data-driven applications and AI/ML approaches. Object storage systems, designed for scalable and cost-effective data management, provide an ideal solution for housing massive language models and large datasets.

This solution uses S3 compatible object storage for storing Llama 2 model (an open-source pre-trained and fine-tuned large language model from Meta) and provides storage for additional dataset and artifacts.

Solution architecture

High-level architecture

Dell APEX Cloud Platform for Red Hat OpenShift is designed collaboratively with Red Hat to optimize and extend OpenShift deployments on-premises with an integrated operational experience.

This turnkey platform provides:

- Deep integrations and intelligent automation between layers of Dell and OpenShift technology stacks, accelerating time-to-value and eliminating the complexity of management using different tools in disparate portals.
- A bare metal architecture that delivers the performance, security, and linear scalability required to meet even the most stringent SLAs.
- Intrinsic multi-layer security, rapid availability of patches and updates, and centralized OpenShift governance to help enterprises maintain a strong security posture.

Figure 3 and Figure 4 show a high-level overview of the physical and logical architecture, including the hardware and software layers. For more information about the solution architecture, see the Design Guide—Implementing a Digital Assistant with Red Hat OpenShift AI on Dell APEX Cloud Platform.

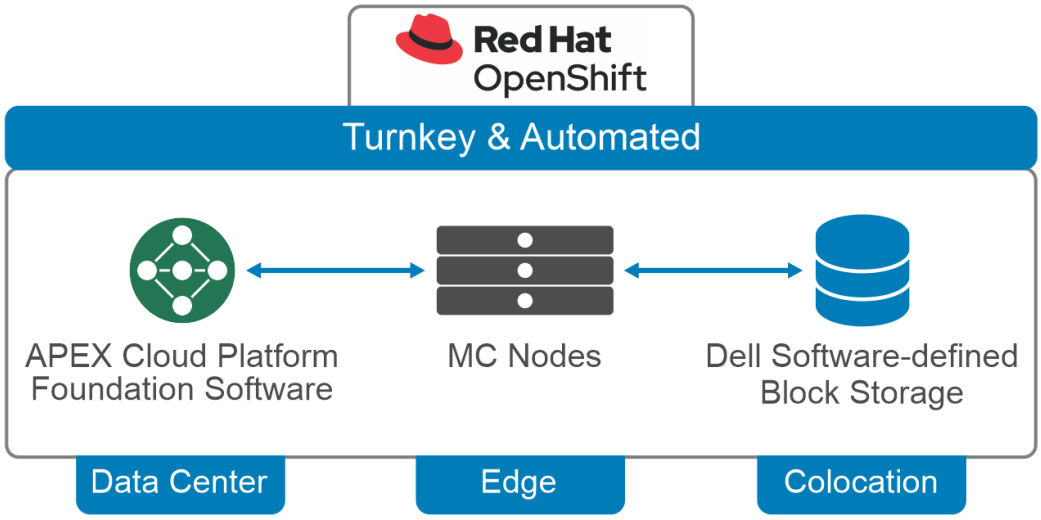

Figure 3. Dell APEX Cloud Platform for Red Hat OpenShift architecture overview

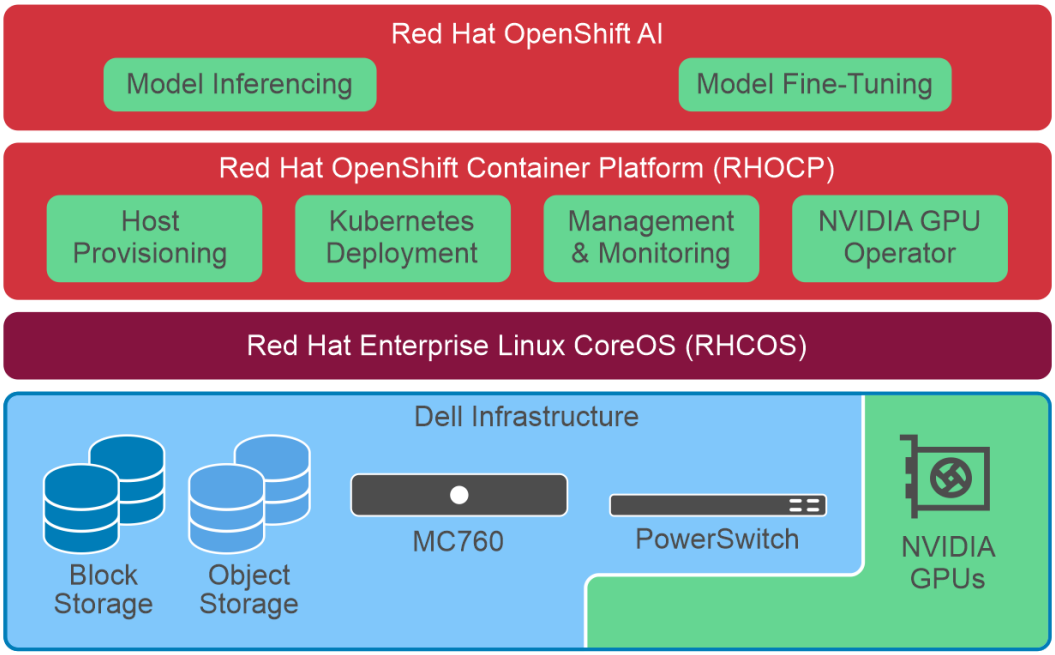

Figure 4. Red Hat OpenShift AI on Dell APEX Cloud Platform for Red Hat OpenShift

Model serving

Llama 2 deployment

Deploying a LLM is a multi-step process which can be challenging. A robust AI/ML platform requires the right combination of hardware and software.

Dell APEX Cloud Platform for Red Hat OpenShift reduces this complexity by providing an AI/ML solution for data scientists and data engineers to seamlessly deploy a LLM model.

In this exercise, we deployed a LLM based digital assistant on Dell APEX Cloud Platform for Red Hat OpenShift that can answer user queries related to domain specific documents. As text-based searches are limited in obtaining the right data, a digital

assistant can help retrieve more accurate and relevant results using semantic search and natural language processing.

The following list describes the AI components used in this exercise.

- **Llama 2:** A second generation, open-source, pre-trained, and fine-tuned LLM ranging from 7B to 70B parameters. It can be used to build chatbots, language generation, and other AI-powered tools.
- **RAG:** An AI framework that combines pre-trained language models with a retrieval mechanism. It acts as a bridge between the language models and a repository with large amounts of data, helping LLMs provide better and more accurate answers.
- **LangChain:** A framework for developing applications powered by language models. It is developed to simplify the process of building LLM powered applications by providing an abstracted standard interface that makes it easier to interact with different language models, including Llama 2.
- **Gradio:** Gradio is an open-source Python library that enables incredibly fast development and prototyping of ML web applications with user interfaces. It provides a simple and intuitive API which is compatible with all Python programs and libraries. Gradio provides a variety of options to customize various elements of the user interface (UI).
- **Redis:** Redis is a popular in-memory data structure store. One of the features of the Redis database is the ability to store embeddings with metadata for LLMs to use. Redis vector database is an excellent choice for applications that need to store and search vector data quickly and efficiently.

Combining these AI technologies with Dell APEX Cloud Platform for Red Hat OpenShift and Red Hat OpenShift AI offers organizations an efficient way to deploy an integrated platform for AI/ML workloads. This solution mitigates prior complexities and enables data scientists to efficiently develop, deploy, and manage the full life cycle of data science and ML workloads.

Conclusion

As more data becomes available and algorithms become more sophisticated, AI and ML technologies continue to advance. To remain competitive in this digital age, organizations must embrace new technologies while maintaining operational and cost efficiency.

Dell Technologies and Red Hat have teamed up to deliver customers a full-stack AI/ML solution built on Dell APEX Cloud Platform for Red Hat OpenShift with Red Hat OpenShift AI.

The combined effort between Dell Technologies and Red Hat introduces a powerful solution that brings reliability, performance, and scalability for AI/ML workloads. The result is a dynamic ecosystem that empowers organizations to exploit the full potential of their data, optimize ML workloads, and drive transformative insight, enabling businesses to make data-driven decisions quickly.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email.

References

Dell Technologies documentation

The following Dell Technologies documentation provides additional and relevant information. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- Dell Validated Design—Implementing a Digital Assistant on Dell APEX Cloud Platform for Red Hat OpenShift
- Dell Technologies Solutions Info Hub for APEX Cloud Platforms
- Dell APEX Cloud Platform for Red Hat OpenShift

Red Hat documentation

The following Red Hat documentation provides additional and relevant information:

- Red Hat OpenShift AI
- Red Hat OpenShift Container Platform