Introduction to Red Hat Enterprise Linux AI 1.1 on Dell PowerEdge GPU Enabled Servers

Enabled by Dell PowerEdge Servers

September 2024

H20138

White Paper

Abstract

This white paper provides an overview of the features and capabilities of Red Hat Enterprise Linux AI on Dell GPU Enabled Servers.

Dell Technologies AI Solutions

Dell
Reference Design

Copyright

The information in this publication is provided as is. Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Dell Inc. believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Contents

Executive summary

Overview

Red Hat Enterprise Linux AI (RHEL AI) serves as a foundation model platform, seamlessly facilitating the development, testing, training, and inferencing of generative AI (GenAI) models for enterprise applications.

RHEL AI brings together:

- The Granite family of open source Apache 2.0-licensed large language models (LLMs) with complete transparency on training datasets.
- InstructLab model alignment tools, which open the world of community developed LLMs to a wide range of users, that offer fine-tuning of models via the Large-scale Alignment for ChatBots (LAB) technique.
- A bootable image of RHEL including popular AI libraries such as PyTorch and hardware optimized inference for NVIDIA, Intel, and AMD.
- Enterprise-grade technical support and model intellectual property indemnification provided by Red Hat.

RHEL AI allows portability across hybrid cloud environments and makes it possible to scale your AI workflows either on Red Hat OpenShift AI or inference them on RHEL AI.

This white paper presents a reference implementation for RHEL AI on Dell PowerEdge servers. This paper also provides a brief overview of Red Hat Enterprise.

Audience

This document is intended for data scientists, data engineers, AI developers, developers, system administrators, and architects. Experience with RHEL is recommended, but not required.

Revisions

Date	Part number/ revision	Description
September 2024	H20138	Initial release

We value your feedback

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

Contributors: Jacob Kirby (Dell), Deepak Aakula (Dell), Matthew Miller (Red Hat), Jeremy Eder (Red Hat)

Note: For links to other documentation for this topic, see the Specialty Servers Info Hub.

Solution overview

Business challenges

- **Resource Constraints**: Many organizations lack the necessary hardware and user-generated data resources to undertake resource-intensive training or fine-tuning of LLMs. RHEL AI provides a comprehensive suite of fine-tuning utilities (based on LAB) that minimizes dependence on costly human annotations and proprietary models, while allowing simplified adoption and updates using a composed appliance based model that can run on a single server.
- **Freedom and Flexibility**: Allows organizations to quickly get started with generative AI by bringing a solution closer to the data assets, with tooling that is adoptable by both professional data scientists and developers.
- **Compliance and Regulatory Issues**: Many countries have laws that require data to be stored within their borders, thus limiting their capability to use distributed services that are based in other countries. RHEL AI provides a way to fine-tune models within the borders/boundaries of your controlled environment.
- **Security**: Many organizations worry about the security and control of their data. RHEL AI uses Red Hat Enterprise Linux Image Model, which allows users to deploy and manage RHEL AI capabilities with a secure bootc container image. For more information see: https://www.redhat.com/en/technologies/linux-platforms/enterprise-linux/image-mode.

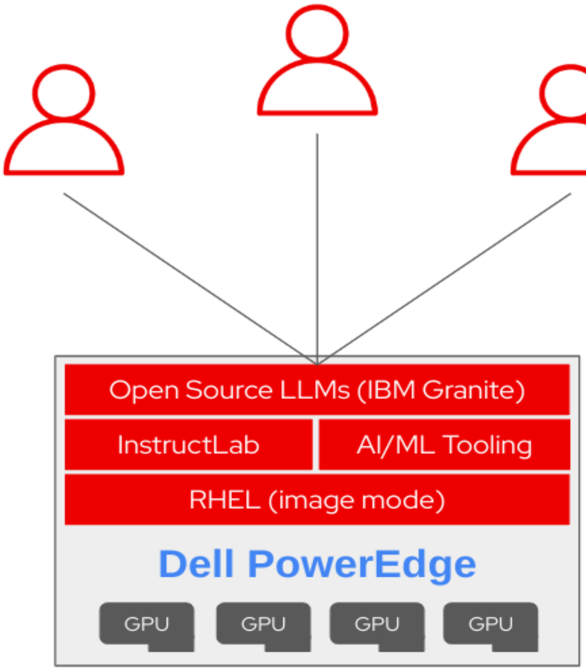

Figure 1. RHEL AI Appliance

Business benefits

- **Open Source Innovation**: RHEL AI harnesses open-source models and tools, enabling businesses to capitalize on community-driven innovation and collaboration.

- **Optimized Performance**: The platform incorporates a streamlined system image that benefits from hardware acceleration, ensuring optimal performance across diverse hardware platforms.
- **Enterprise Support**: RHEL AI offers enterprise-grade technical support, ensuring that businesses have the assistance they need to deploy and maintain AI solutions effectively.

Solution benefits

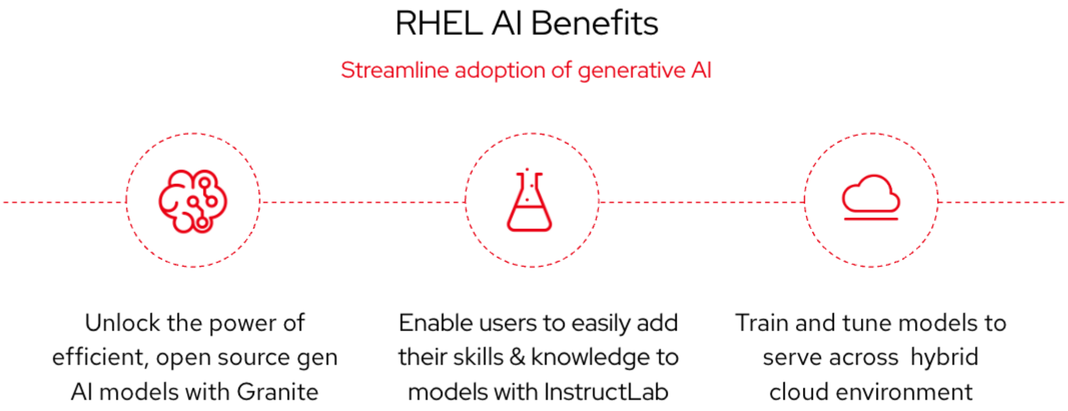
RHEL AI Benefits
Streamline adoption of generative AI
Unlock the power of efficient, open source gen AI models with Granite
Enable users to easily add their skills & knowledge to models with InstructLab
Train and tune models to serve across hybrid cloud environment

Figure 2. RHEL AI benefits

Solution design

Technology overview

RHEL AI

Includes the following hardware and software components:

- Dell PowerEdge Servers
- Industry-leading NVIDIA GPUs for AI/ML
- Red Hat RHEL AI bootable container image based on RHEL image mode

Dell PowerEdge servers with GPU

RHEL AI has been validated on the following servers:

- PowerEdge XE9680 with 8 x Nvidia H100 SXM GPU
- PowerEdge 760xa with 4 x Nvidia H100 PCIe GPU with NVLink
- PowerEdge 760xa with 4 x Nvidia L40S PCIe GPU with NVLink

Figure 3. PowerEdge XE9680 Server

Figure 4. PowerEdge R760xa Server

NVIDIA GPUs for AI/ML

- Nvidia H100 SXM GPU
- Nvidia H100 PCIe GPU
- Nvidia L40S GPU

Red Hat RHEL AI

RHEL AI is a foundation model platform that combines open-source LLMs, enhanced tooling, and a community-driven approach to generative AI model development.

Deployment models

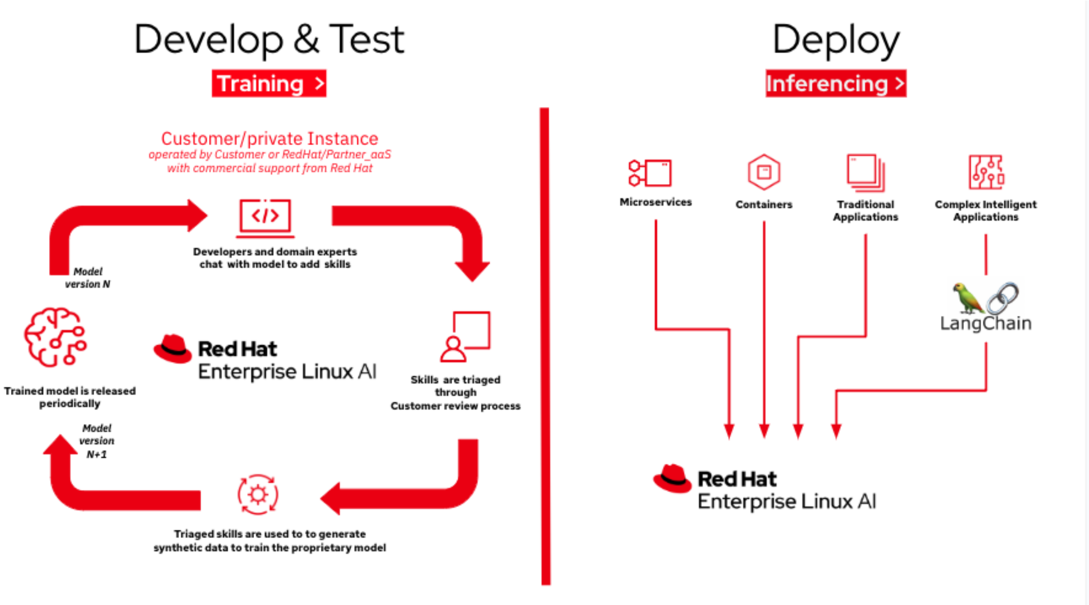
Develop & Test
Training >
Customer/private Instance
operated by Customer or RedHat/Partner_aaS
with commercial support from Red Hat
Developers and domain experts chat with model to add skills
Model version N
Red Hat Enterprise Linux AI
Skills are triaged through Customer review process
Trained model is released periodically
Model version N+1
Triaged skills are used to to generate synthetic data to train the proprietary model
Deploy
Inferencing >
Microservices
Containers
Traditional Applications
Complex Intelligent Applications
LangChain
Red Hat Enterprise Linux AI

Figure 5. AI Lifecycle using Red Hat Enterprise Linux AI

Hardware design

The RHEL AI server is self-contained in a single PowerEdge server including the GPUs(s).

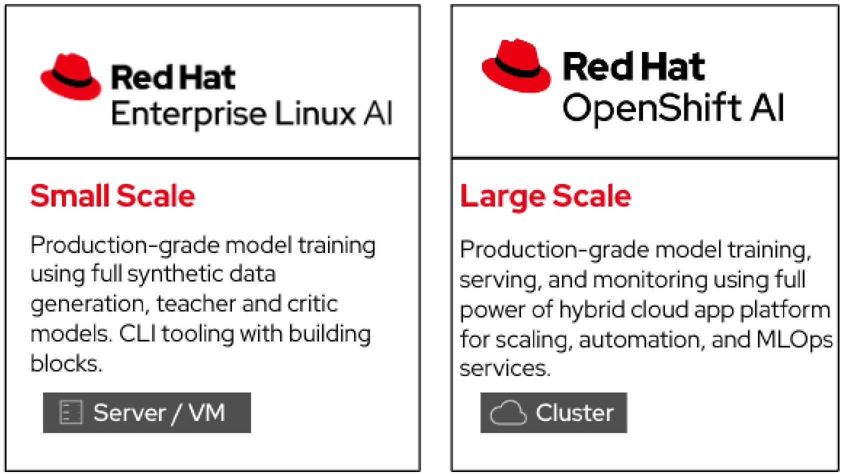
Red Hat Enterprise Linux AI
Small Scale
Production-grade model training using full synthetic data generation, teacher and critic models. CLI tooling with building blocks.
Server / VM
Red Hat OpenShift AI
Large Scale
Production-grade model training, serving, and monitoring using full power of hybrid cloud app platform for scaling, automation, and MLOps services.
Cluster

Figure 6. Small and large scale Red Hat Deployment options

Storage

RHEL AI leverages the local storage of the PowerEdge server on which it runs. Local disks are recommended to be solid-state storage.

Networking

Each RHEL AI server is a standalone server, requiring no specialized networking beyond standard network access for client access, and access to external storage and other AI infrastructure.

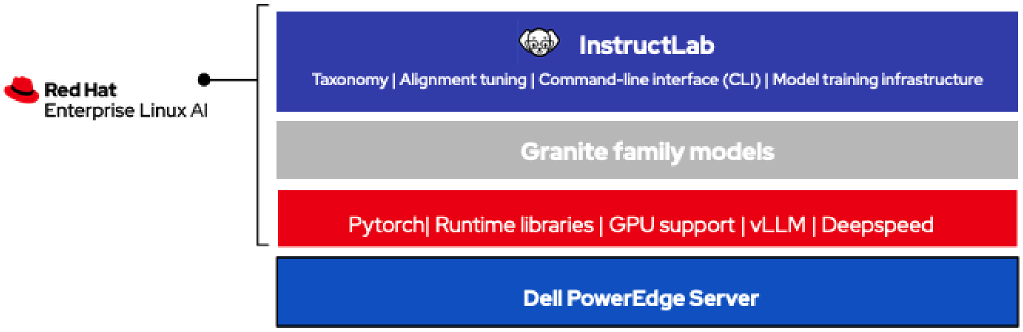
Red Hat Enterprise Linux AI
InstructLab
Taxonomy | Alignment tuning | Command-line interface (CLI) | Model training infrastructure
Granite family models
Pytorch| Runtime libraries | GPU support | vLLM | Deepspeed
Dell PowerEdge Server

Figure 7. Red Hat Enterprise Linux AI software stack

Open-source IBM Granite models

The IBM Granite family of artificial intelligence (AI) models include:

InstructLab

InstructLab, an open-source project created by IBM Research and Red Hat, based on the aforementioned LAB method, aims to enhance LLMs used in generative AI applications by making fine-tuning of foundation models straight-forward and accessible.

It leverages three components:

- **Taxonomy-based data representation**: Humans add and curate knowledge and skill data to the taxonomy for model training.
- **Large-scale synthetic data generation**: A teacher model creates many new examples based on the manually input seed training data. A critic model filters the synthetic data for accuracy and safety.
- **Iterative alignment tuning**: The LLM is retrained using synthetic data.

InstructLab provides a cost-effective way to enhance pre-trained large language models to fit specific business requirements.

RHEL in image mode

Image mode for RHEL offers a simplified approach to managing the RHEL operating system. It allows administrators to build a bootable container image in the same manner that a container image is built. Container images are also used for image updates. This brings operating system builds, updates, and maintenance into the same CI/CD operational workflows as an application.

Implementation guidance

Refer to Red Hat installation documentation here.

Security considerations

Deploy and use the hardware and software recommended in this solution securely. Follow recommendations from Dell Technologies and the other vendors cited in this technical white paper. For additional information, refer to the Dell Technologies Security and Trust Center at https://www.dell.com/en-us/dt/about-us/security-and-trust-center/index.htm.

Conclusion

RHEL is the world’s leading enterprise Linux platform, certified on hundreds of clouds and with thousands of hardware and software vendors. With the technological foundation of Linux, containers, and automation, Red Hat’s open hybrid cloud strategy gives you the flexibility to run your AI applications anywhere you need them.

RHEL AI, Granite, and InstructLab further deliver on this vision, breaking down the cost and resource barriers to experimenting with and building AI models while providing the tools, data, and concepts needed to fuel the next wave of intelligent workloads.

References

Dell Technologies documentation

The following links provide additional information about the solution design and components used in this design. Other reference links are provided in context in the body of the document.

- Dell Technologies Info Hub for AI Solutions

Red Hat Enterprise Linux AI documentation

- Red Hat Enterprise Linux AI documentation
- Red Hat Enterprise Linux AI | Red Hat Developer - Overview and resources for RHEL AI
- What is RHEL AI? A guide to the open source way for doing AI - Detailed guide on RHEL AI and its open-source approach
- Open Source Powered AI/ML for the Hybrid Cloud | Red Hat Developer - Information on AI/ML tools and resources for hybrid cloud environments
- RHEL AI · GitHub - GitHub repository for RHEL AI projects

Other resources

- arxiv logo > cs > arXiv:2403.0108: LAB: Large-Scale Alignment for ChatBots
- Hugging Face > InstructLab
- InstructLab - A new community-based approach to build truly open-source LLMs