Integrate and simplify production AI with Intel and Red Hat

Accelerate the delivery of AI-enabled applications and reduce time to market by empowering data scientists with a choice of applications and services from a vast partner ecosystem.

Experiment, develop, publish, deploy, and maintain AI/ML applications—without worrying about infrastructure or platform lock-in.

Employ powerful Intel software, such as Intel oneAPI AI Analytics Toolkit, Convrg.io, and OpenVINO Pro for Enterprise—all from within Red Hat OpenShift.

Operate and deploy across diverse hardware on demand, using Intel Xeon processors, Intel Habana Gaudi deep-learning processors, and Intel graphics processing units (GPUs).

Taking models from concept to production

Myriad organizations are embracing artificial intelligence and machine learning (AI/ML) to improve the customer experience, make employees more productive, accelerate innovation, and develop new products with increased speed. Generative AI adoption is increasingly mainstream, with developers across all domains and industries exploring ways to benefit from these foundational models. Despite these attractive benefits and significant industry-wide investments, a high AI/ML project failure rate remains, with many experiencing real challenges moving projects from concept to production.[1] These challenges are understandable, as data science encompasses many tasks—data acquisition, model building, development processes, and continuous integration and continuous delivery (CI/CD), to name a few.

With more than 25 years of working together, Intel® and Red Hat can help organizations face these challenges directly and rapidly operationalize their AI/ML models. Red Hat® OpenShift® AI provides an AI-focused portfolio with tools that span the entire life cycle of AI/ML experiments and models. In collaboration with Red Hat, Intel furnishes proven, optimized libraries and toolkits, including Intel oneAPI AI Analytics Toolkit (AI Kit) and OpenVINO Pro for Enterprise. These products equip data scientists with the latest tools and technologies optimized for Intel architecture-based hardware, including Intel Xeon processors, Habana Gaudi for dedicated deep-learning training and inference, Intel Data Center GPU Flex Series for AI inference, and Intel Data Center GPU Max Series for AI inferencing and training.

Simplifying AI/ML with Red Hat OpenShift AI and Intel

Data scientists increasingly depend on open source technologies for AI/ML solutions. OpenShift AI integrates these technologies into an enterprise-grade hybrid AI and ML operations (MLOps) platform. With an AI-focused portfolio, OpenShift AI provides tools across the full life cycle of AI/ML experiments and models. The resulting platform is a consistent, unified, and scalable foundation for IT operations (ITOps) leaders while providing data scientists and developers with tooling and access to accelerators to capture AI innovations.

OpenShift AI combines proven AI capabilities and Red Hat OpenShift—an open source application platform for accelerating the development and delivery of cloud-native applications. As an add-on to Red Hat OpenShift, OpenShift AI offers a robust environment for building and deploying models into production, combining the self-service environment that data scientists and developers want with the deployment confidence enterprise IT organizations demand. Data scientists can experiment with a choice of tools, collaborate, and accelerate time to market—all within a common platform. Teams can collaborate with less friction and deliver intelligent applications to the market more efficiently, ultimately delivering more value for the business. Coupled with Intel software and hardware offerings, Red Hat OpenShift AI brings key benefits to data scientists and development teams, including:

1 *According to 2 recent Gartner reports, 85% of AI/ML projects fail to deliver and only 53% of projects make it from prototypes to production. https://www.infoworld.com/article/3639028/why-ai-investments-fail-to-deliver.html.*

Not all AI/ML applications require GPU acceleration. Intel estimates that 70% of datacenter inferencing runs on Intel Xeon.

Intel's software strategy embraces an open ecosystem with open platforms and open source software, libraries, and application programming interfaces (APIs). Intel's oneAPI industry initiative empowers developers to innovate with speed and agility, pushing the limits of performance while ensuring security. By offering choice and trust, Intel software creates unique competitive advantages and disruptive growth potential for our customers, the ecosystem, and the world.

- Eliminating complex ITOps setup tasks.
- Managing software life cycles efficiently through a choice of a managed cloud service or traditional software deployment methods.
- Providing specialized components and integrated independent software vendor (ISV) support within the user interface (UI).
- Allowing development and deployment on any environment across the datacenter, the edge, and public clouds.
- Serving models and automating data science pipelines at scale to simplify MLOps processes.

Intel AI tools and transparent hardware acceleration

Intel technologies are integrated and optimized with open source AI/ML tools and Red Hat OpenShift AI to provide a simplified experience for data scientists, developers, and operations personnel.

- **Intel AI Analytics Toolkit.** Accessible directly from within a Jupyter Notebook in the Red Hat OpenShift AI environment, Intel AI Analytics Toolkit powered by oneAPI is optimized for maximum performance on Intel architecture-based central processing units (CPUs) and auxiliary processing units (XPUs) and is available as a Red Hat OpenShift operator.
- **OpenVINO Pro for Enterprise.** Featuring a collection of pretrained models that perform various inference tasks, OpenVINO Pro for Enterprise maximizes the performance of computer workloads across Intel architecture-based hardware and supporting AI applications.
- **Habana Gaudi series.** With the power of Intel technology, Habana Gaudi delivers unparalleled performance and efficiency, providing rapid, more accurate AI computations for ML algorithms, neural networks, or data analytics.
- **Cnvrg.io.** Cnvrg.io simplifies the entire MLOps life cycle, automating MLOps processes and allowing you to smoothly integrate your ML workflows while efficiently managing, organizing, and scaling your projects.
- **Intel XPUs for performance and accelerated inference.** With support for Intel Xeon processors, Intel Habana deep-learning processors, and Intel GPUs, data scientists can select an architecture well-suited for a particular AI solution and reuse their code—without rewriting it—for another architecture and platform.

redhat.com
#628003_1223