Solution Brief

Streamline AI Adoption and Deployment Using Intel Enterprise AI with Red Hat® OpenShift® AI

Red Hat and Intel offer curated, validated and integrated hardware and software elements for enterprise AI. Taken together, they reduce the complexity of selecting and integrating solution components, helping customers attain faster time to value, lower cost and less risk in their mainstream infrastructures, with enterprise customer support.

Authors:

Abhay Chitral, Intel
Karl Eklund, Red Hat
Julie Fleischer, Intel
Michael Hrivnak, Red Hat
Sridhar Kayathi, Intel
Raghu Moorthy, Intel

Mainstream business strategies are now routinely informed by AI-powered insights, to match supply to demand, predict market changes and improve product and service offerings. Indeed, 79% of corporate strategists agree that AI is critical to success for their business in 2024.[1] In pursuit of those initiatives, the global AI market is growing at a CAGR of 37% through 2030,[2] demonstrating the widespread acknowledgment of its strategic importance.

79% of corporate strategists **say AI is critical to success in 2024**[1]

37% CAGR, 2023-2030 **growth is underway in the global AI market**[2]

Even as AI refactors operations for businesses across industries, generative AI (GenAI) is supercharging both organizational capabilities and customer demand. Large language models (LLMs) that incorporate billions or trillions of parameters are becoming more capable at assisting or even emulating human thinking. As algorithms become more effective workforce multipliers, the use of GenAI is emerging as a competitive imperative. While just 10% of organizations launched AI solutions to production in 2023,[3] Gartner predicts that 80% of enterprises will use GenAI by 2026.[4]

10% of organizations **launched GenAI solutions to production in 2023**[3]

80% of enterprises **will use GenAI by 2026**[4]

Despite that consensus and investment, AI benefits have proved difficult to realize. Successfully implementing AI as a business driver requires evolution of a company's technology landscape, which must be directed using novel skill sets that most organizations do not have in-house, and which are in short market supply. The lack of longstanding expertise can make it difficult to navigate the overabundance of hardware and software options available to AI project teams. The complexity of this transition often drives up the cost of GenAI initiatives, jeopardizing their progress.

Moreover, the immense scope of GenAI's potential value creates rapid development and churn in the technologies and tools available to drive change. That moving target contributes to the reality that 70% to 80% of AI projects never make it to production.[5] New processes and pipelines are needed to graft AI's potential onto existing business operations to enable a data-driven future, and 46% of experts cite infrastructure as the biggest challenge in productizing LLMs.[3]

70-80% of AI projects
never make it into production[5]

46% of experts
say infrastructure is the biggest challenge in productionalizing LLMs[3]

The innovation to overcome these obstacles is exhibited in a multitude of discrete open source projects both large and small, from the most popular deep learning frameworks to obscure domain solutions. As always, however, the value of this innovation must be realized in deployment. Together with this undeniable wealth also comes complexity, and the task of choosing the right components often falls to the customer.

While open source software ecosystems offer interoperability advantages over proprietary ones, architects face too many choices. As they look ahead to making AI a cornerstone of their technology and business strategies, decision makers must also consider the full lifecycle implications of the building blocks they choose.

For example, community distributions of important solution components may fall short in terms of providing targeted integration expertise as well as meeting requirements for ongoing enterprise-class support.

To help customers overcome the challenges of a complex and dynamic ecosystem, Red Hat and Intel offer leadership transforming open technologies into solutions. The companies each draw on their decades of open source expertise, separately and in collaboration, to help customers bring their AI visions and projects from concept to production.

Out-of-the-box enterprise AI with Red Hat and Intel

To help implementation teams tailor optimized technology stacks for their solutions, Red Hat and Intel each offer a curated set of optimized components that they have validated and integrated to work together. To increase that value even further, they have combined those ecosystems to work smoothly in combination with each other, enabling high-quality AI solutions with superior quality, time to market and cost efficiency:

- Red Hat® OpenShift® AI is an enterprise-ready AI platform, built for cloud-era innovation on top of Red Hat OpenShift.
- Intel enterprise AI incorporates Intel AI software and selected hardware, Operators and cloud instances, unified under an umbrella of enterprise engagement.

Together, this portfolio — illustrated in Figure 1 — gives customers a growing set of curated, cutting-edge technologies as well as structured guidance about how to tailor them to individual project needs. The shared design philosophy by both Red Hat and Intel embraces the need to be comprehensive without being prescriptive. This approach serves to give architects and other decision makers a strong foundation that meets diverse project needs, combined with the ability to help them innovate without limit.

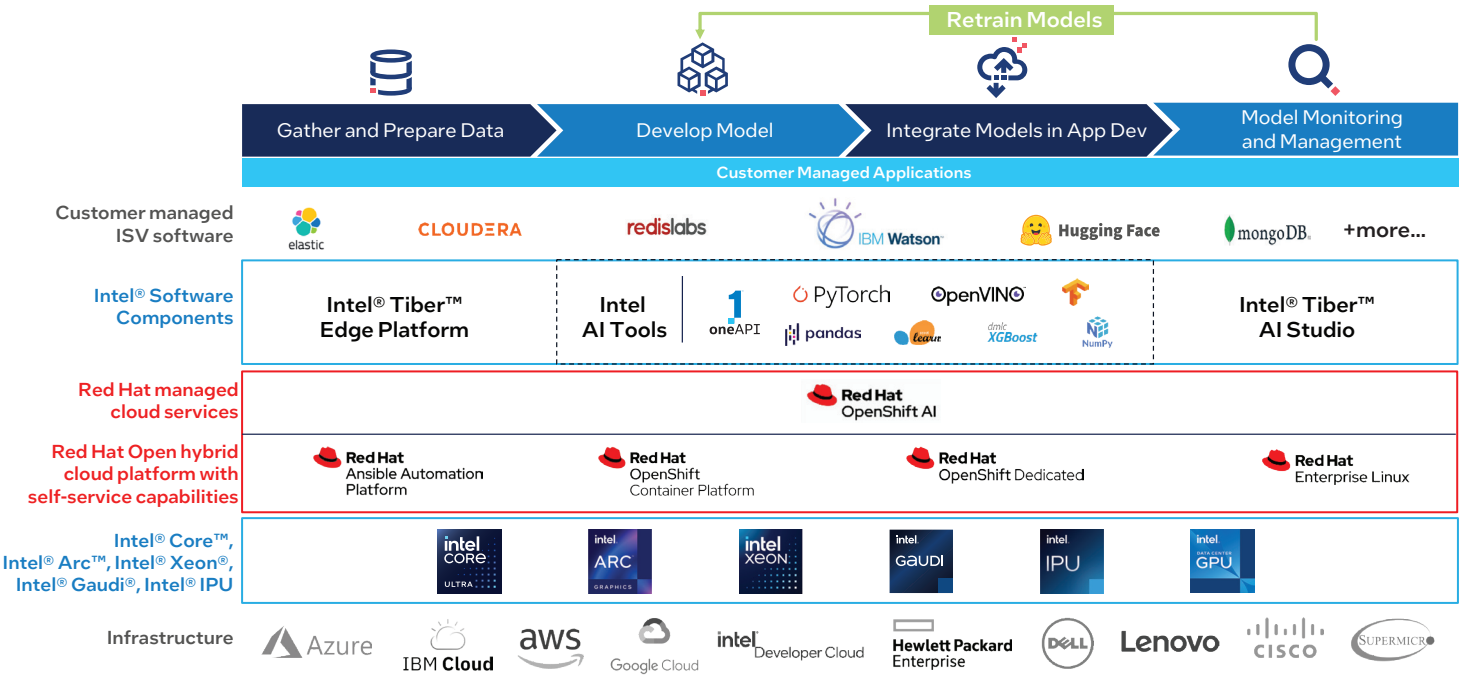

Figure 1. Joint ecosystem: Red Hat® OpenShift® AI and Intel enterprise AI.

At the foundation of the technology stack is the range of standards-based Intel architecture platforms, including both conventional hardware and IaaS from leading cloud service providers (CSPs). These infrastructure options enable interoperation by customer workloads among CPUs, GPUs and Accelerators, for optimized results across key performance indicators (KPIs) such as time- and cost-to-train, as well as inference throughput and latency, for varied usages and requirements.

Red Hat OpenShift and Red Hat OpenShift AI sit above the infrastructure, as the software foundation for AI solutions. Red Hat OpenShift is an enterprise Kubernetes platform for container development, deployment and orchestration. Red Hat OpenShift AI is built on OpenShift using open source software components, providing a platform where data scientists and developers work together to build solutions. It includes libraries and pre-built models to enable and accelerate development, right out of the box.

Intel AI software comprises a comprehensive set of open source components, validated and verified to work together, as the basis for building and operationalizing AI applications. The software packages are optimized for Intel architecture using oneAPI, providing an open programming model and application enablement for Accelerators and other features and capabilities across Intel platforms. These tools span the entire AI pipeline, including data preparation, training, training, fine-tuning and inference. At the top of the stack is the broader open source ecosystem, which includes additional optimizations and contributions from Red Hat and Intel.

Customers now have structured means to navigate the open source landscape to advance their AI objectives, selecting their choice of components from the global community. Their architectures transparently inherit world-class optimizations and assurances of interoperability from Red Hat and Intel.

Artificial intelligence platform: Red Hat® OpenShift® AI

Red Hat OpenShift AI supports the full lifecycle of AI/ML experiments and models, on-premises and in the public cloud. Illustrated in Figure 2, it provides a comprehensive resource for teams to work with their choice of tools, collaborate on a common platform and bring solutions to market quickly and successfully. Red Hat OpenShift AI is available as an add-on to Red Hat OpenShift, either as a fully managed cloud service or as a self-managed software product.

This trusted foundation, built on open architecture, is engineered to streamline the process of building AI solutions and bringing them online, shrinking the gaps between data science and DevOps. There is no prescriptive toolchain; data scientists can use their familiar tools as well as draw from a growing partner ecosystem. Self-service infrastructure enables data teams to add software or spin up resources as needed, without waiting for IT. This helps eliminate the common problem of rogue accounts that can create security exposures and other headaches for the organization as a whole.

Data scientists use JupyterLab to conduct data exploration and develop models using provided notebooks or their own, with access to optimized libraries and frameworks including TensorFlow and PyTorch. Organizations can create repeatable pipelines to formalize and streamline model training and validation, integrated with existing DevOps practices for secure and effective delivery. Models can be deployed across any cloud, with full centralized management and performance monitoring. Solutions draw on any combination of open source and ISV software.

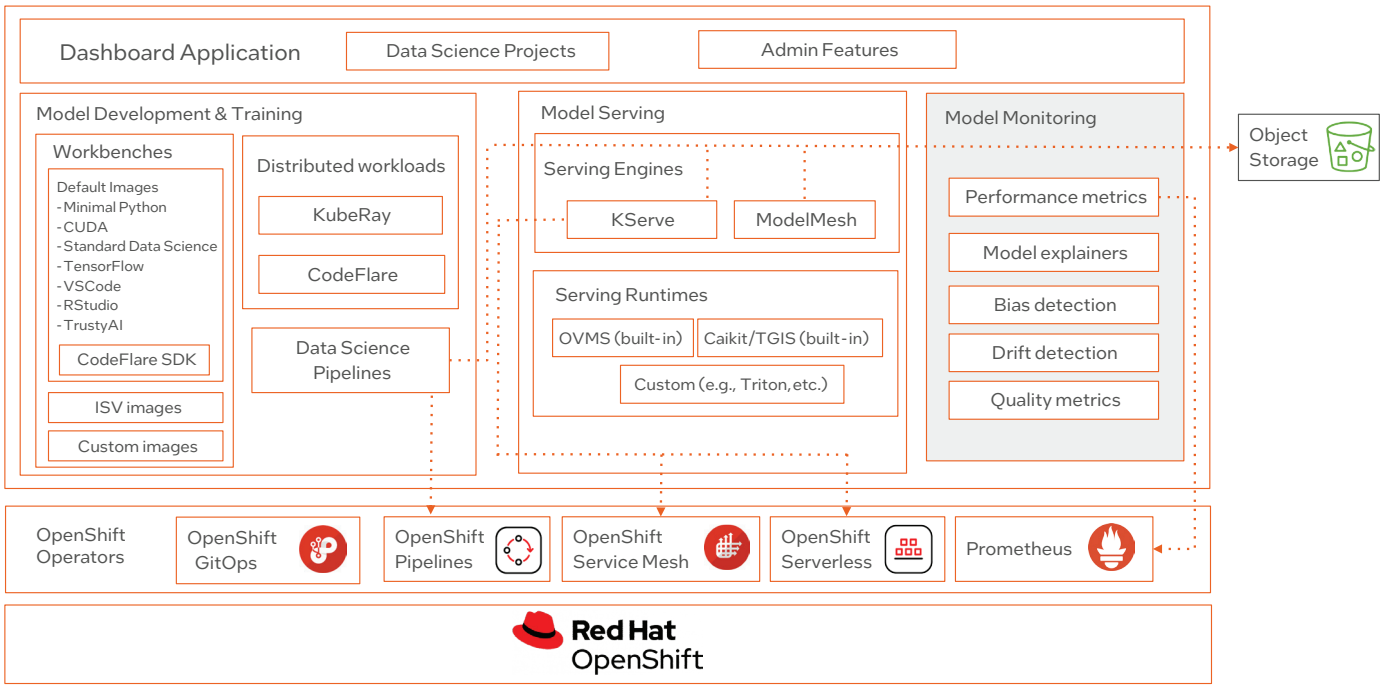

Figure 2. Red Hat® OpenShift® AI: A common platform for diverse teams and roles.

Red Hat OpenShift AI extends the DevOps pipeline to include machine learning. It lets model artifacts be treated in the same way as code artifacts, enabling rigorous, unified workflows from development to production. Because both data scientists and developers work under a common set of DevOps principles, handoffs have less friction. This means that developers can integrate AI capabilities into their applications with more confidence, bringing solutions into production more quickly and smoothly, with less risk. Platform workflow includes the following components:

- Starburst Galaxy, built on the open source Trino SQL engine, enables teams to make fast, easy queries against diverse data sets and draw insights from them wherever they are, across hybrid cloud infrastructures.
- Anaconda provides an extensive repository of curated data science packages for use in Jupyter projects, with pre-built Jupyter images available directly from the Red Hat OpenShift AI dashboard.
- Pachyderm enhances data governance in the pipeline-creation process, with guaranteed data lineage provided by automatic data versioning.

Red Hat OpenShift is the secure, proven, supported hybrid cloud platform at the foundation of Red Hat OpenShift AI. Built on top of Kubernetes, OpenShift is enhanced for the enterprise with hardening and subscription-based technical support to help optimize security, reliability and ease of integration. OpenShift Container Platform adds further tooling for functionality that includes cluster services, workload management, code development and developer productivity.

Software and hardware ecosystem: Intel enterprise AI

Intel's offerings in enterprise AI are based on an open, modular platform that aims to facilitate and enable the development of flexible, scalable AI systems. They harness open source innovations from across the ecosystem and consist of an end-to-end suite of AI software and hardware ingredients, illustrated in Figure 3, optimized and integrated to work together. This engineering work helps ensure a fast, smooth development and deployment path with high performance and solution quality.

The platform brings together Intel platform expertise with enablement across the ecosystem, including certified solutions on partner software, systems and instances for data science, machine learning and AI. Intel enterprise AI is also the medium for enterprise engagement with customer data science and DevOps teams to proliferate success based on deep understanding of usages and workloads, from the client and edge to the data center and cloud.

Integration between Intel enterprise AI and Red Hat OpenShift AI benefits from longstanding collaboration between the companies for co-engineering and joint enablement. Engineering teams from Red Hat and Intel work together from early in the product development cycle to ensure that their products run together smoothly with mutually enhancing features for production AI.

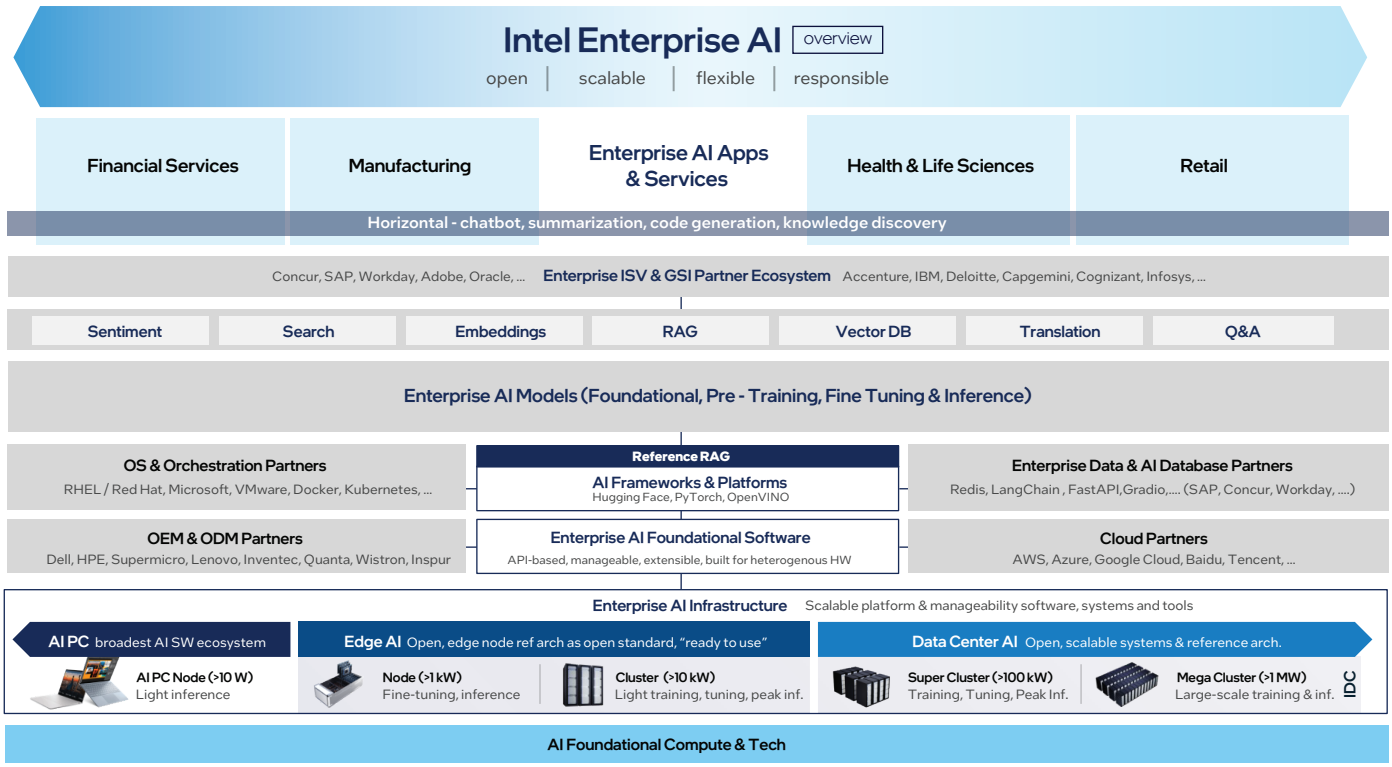

Figure 3. Intel Enterprise AI components.

Intel enterprise AI software stack

Intel follows a software-first strategy for AI, built from the ground up with developers in mind. By eliminating the need to code for specific hardware architectures, this approach speeds up development cycles to accelerate time to market and improve return on investment, with higher solution quality and less risk.

oneAPI is the open, cross-architecture specification and implementation at the heart of the Intel ecosystem that replaces proprietary GPU-focused approaches for AI. The two primary oneAPI toolkits used in Intel enterprise AI are OpenVINO™ and Intel AI Tools.

OpenVINO is an open source toolkit powered by oneAPI for optimizing and deploying deep learning models with fewer lines of code than would otherwise be possible, with a "write-once, deploy-everywhere" model across Intel platforms and a post-training optimization tool. It allows models to be converted from different frameworks such as TensorFlow and PyTorch into an Intermediate Representation format that can be easily deployed and optimized on a wide variety of hardware, including CPUs, discrete or integrated GPUs and FPGAs.

Intel AI Tools, shown in Figure 4, is a oneAPI toolkit that helps speed up time to market and accelerate end-to-end machine learning and data science pipelines with optimized deep learning frameworks and high-performing Python libraries. Intel AI Tools includes support for Intel-optimized implementations of PyTorch and TensorFlow as well as classic machine learning packages.

Organizations can accelerate AI development by building and tuning oneAPI multi-architecture applications on Intel® Tiber™ Developer Cloud. This coding sandbox provides the latest optimized oneAPI and AI tools, enabling workloads to be tested across Intel CPUs and GPUs. It includes pre-release Intel platforms and associated Intel-optimized software stacks, including the latest machine learning toolkits from Intel and libraries hosted on Intel Developer Cloud. No hardware installation, software download or configuration is necessary.

Intel® Tiber™ Edge Platform offers a streamlined and scalable path to build, deploy and manage edge AI and in-demand edge use cases. It dramatically simplifies the implementation and management of intelligent edge solutions. Solutions built on Intel Tiber Edge Platform are highly flexible and open, allowing transformative edge use cases on your existing infrastructure. They also deliver smart vision, predictive analytics and other demanding solutions without specialized components.

Intel is also a member of Open Platform for Enterprise AI (OPEA), a sandbox project from the Linux Foundation that is working to foster development of composable (modular) building blocks for hardened, scalable GenAI solutions. In particular, OPEA is working on pipelines for retrieval-augmented generation (RAG), which extends the data that models can draw on to generate a response beyond the original training data, making them more up-to-date and effective. Industry collaboration within the OPEA project helps standardize components used for RAG, for open interoperability.

Intel hardware platforms and OpenShift AI

Intel enterprise AI delivers software functionality with deep optimizations across the range of Intel hardware platforms. In contrast with proprietary solutions that tie customers to a specific hardware architecture, Intel enterprise AI is platform-neutral, which enables different hardware capabilities to be utilized depending on cost and speed considerations, which can vary throughout the AI lifecycle.

OpenShift Operators are the structures used by Intel to enable Intel hardware; they provide the Accelerator provisioning and installation. Operators foster repeatability in IT processes, perform ongoing health checks of system components and perform over-the-air software updates. Operators for Intel Accelerators are part of the Red Hat OpenShift Certified Operators, which it offers on a SaaS basis through the Red Hat Ecosystem Catalog.

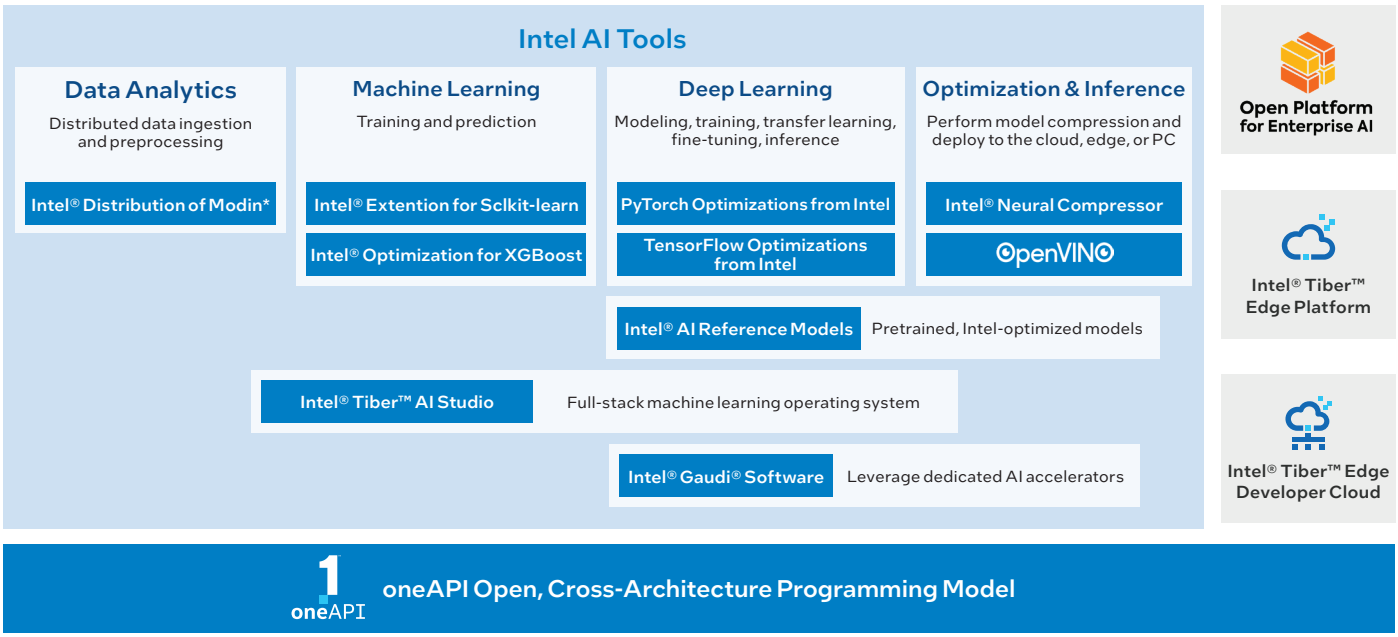

Figure 4. Intel AI Software.

Intel enterprise AI includes the following Operators and other hardware components:

- Intel® Gaudi® AI Accelerators and Intel Gaudi software are built for dedicated, large-scale AI training and inference optimized for compute performance, efficiency, usability and choice. They provide outstanding time- and cost-to-train, efficient model fine-tuning and favorable throughput and latency. The Intel Gaudi Operator for OpenShift AI is located at https://catalog.redhat.com/software/container-stacks/detail/64342b3bcbfbb9a6588ce8dd.
- Intel AI-accelerated CPUs are a cost-effective, familiar option for general-purpose AI and mixed workloads. They enable inference to coexist with other enterprise applications, taking advantage of existing enterprise pipelines for an optimized path from data ingest to inference. Intel® Xeon® processors target AI workloads with the industry's largest set of built-in Accelerators, including Intel Advanced Matrix Extensions (Intel AMX), which dramatically improves throughput and latency. Intel AMX support is enabled directly in OpenShift AI, so no additional Operator is required.
- Intel client / edge GPUs support local execution of LLMs on platforms based on Intel® Core™ Ultra processors, Intel Xeon processors and Intel® Arc™ GPUs. Support for Intel Arc GPUs in OpenShift AI is available upon request.
- Intel® Data Center GPUs support specialized mixed AI workloads with an open programming model. Intel Data Center GPU Max Series is Intel's highest performing, highest density general-purpose discrete GPU, built for HPC and AI. Intel Data Center GPU Flex Series is optimized for media encode/decode stream density and quality as well as visual inference. The Intel GPU Operators for OpenShift AI are located at https://github.com/intel/intel-data-center-gpu-driver-for-openshift.

Intel extends optimizations across all its hardware platforms to a large and growing range of open source projects, creating a comprehensive software ecosystem that data scientists and developers can access and utilize through the combined platform based on Red Hat OpenShift AI and Intel enterprise AI.

Conclusion

Intel enterprise AI with Red Hat OpenShift AI provides a unified, consistent environment for AI solutions, out of the box, where the burden of choosing and integrating together optimized components is shifted away from the customer. Project teams choose the combinations of technologies appropriate to their requirements, without vendor lock-in. This path lets customers apply sparse AI resources to building AI applications instead of focusing on infrastructure, dependencies, optimization and compatibility. Bringing innovation to market using this approach reduces risk, cost and time to market, with enterprise security features and support as solutions progress from testbed to production.

Take the Next Step

Sign up for Intel® Tiber™ Developer Cloud

Try a 60-day trial of Red Hat OpenShift AI

More Information

- Red Hat® and Intel® AI and Machine Learning: The Perfect Combination for Data Scientists
- Essential Tools for Jumpstarting AI Development Projects
- How to Use Intel®-Optimized AI Software in the Cloud
- Speed Up Machine Learning Training on CPUs with AI Tools

Solution provided by:

intel + Red Hat

[1] Skim AI, "10 Enterprise AI Statistics to Know in 2024." https://skimai.com/10-enterprise-ai-stats-to-know-in-2024/.

[2] Grandview Research, "Artificial Intelligence Market Size and Share Report, 2030." https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market.

[3] mlinsider. "The state of Generative AI and Machine Learning at the end of 2023." https://cnvrg.io/wp-content/uploads/2023/11/ML-Insider-Survey_2023_WEB.pdf.

[4] Gartner, October 11, 2023. "Gartner says More than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026." https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026.

[5] Cognilytica, "Top 10 Reasons Why AI Projects Fail." https://www.cognilytica.com/top-10-reasons-why-ai-projects-fail/.

Performance varies by use, configuration and other factors. Learn more at https://www.intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for configuration details.

No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

0524/RM/MESH/356881-001US