The Leader In AI Infrastructure Management

Key Benefits

Centralized AI Infrastructure

Unifies compute resources across environments into a single platform for streamlined management and optimization

Dynamic Allocation of Compute Resources

Adapts GPU resources to fluctuating demands, optimizing for efficiency without additional hardware

Visibility Into Resources, Workloads and Users

Offers comprehensive insights into resource utilization, enabling informed decision-making and strategic allocation

Self-serve Access to ML Tools and Accelerated Compute

Empowers data scientists with on-demand environments, facilitating innovation and reducing development cycle

The Hidden Challenges In AI Infrastructure

Enterprises navigating AI and machine learning face significant challenges in AI infrastructure. Efficiently utilizing and dynamically allocating GPU resources to meet varying workload demands is a key challenge, often leading to resource underutilization or the need for costly hardware expansions. Additionally, aligning the distribution of computing power with strategic business goals and evolving demands is crucial for effective AI workload management. This requires a sophisticated integration of business rules, security, and governance, transforming resource management into a strategic function. Managing GPU resources across dispersed locations and cloud environments further complicates scalability and agility. Moreover, supporting the entire AI lifecycle, including model development and large-scale training, demands an infrastructure that optimizes efficiency while ensuring performance and compliance. To address these challenges enterprises need a solution that provides dynamic allocation, strategic resource management, unified pooling, and comprehensive lifecycle support.

Solution - AI Infrastructure Management

AI Lifecycle Integration

Comprehensive support for all phases of ML development including popular open sources tools, training frameworks and LLM's all delivered as-a-service

Resource Pooling

Pool GPUs from public/private cloud and on-prem simplifying management, enabling seamless access, and enhancing AI operation scalability

Policy Engine

Business rules specifying priorities, quotas, and security, driving dynamic allocation of resources aligned to changing business needs

AI Workload Orchestration

Drives workload execution per Policy Engine leveraging GPU Fractioning to maximize utilization and throughput

GPU Fractioning

Dynamically subdivides GPUs maximizing existing GPU resources, and reducing the need for additional hardware

AI Infrastructure Management represents a transformative approach to managing and optimizing AI resources and operations within an enterprise. It is a comprehensive approach designed to overcome the inherent challenges in traditional AI infrastructure by being dynamic, strategic, and integrally aligned with business objectives. Key features of this infrastructure include:

AI Lifecycle Integration

AI Lifecycle Integration offers as-a-service support for every phase of AI development and deployment. This includes providing development environments for data scientists and managing large-scale training and inference workloads, optimizing efficiency without compromising performance or compliance.

Resource Pooling

Addressing the geographical dispersion of compute resources, resource pooling unifies GPUs from various sources into a single manageable pool. This simplifies resource handling in distributed infrastructures and enhances scalability and agility in AI operations.

Policy Engine

The Policy Engine embodies the integration of sophisticated business rules, security protocols, and governance in AI resource management. It translates enterprise strategies into actionable resource allocation, elevating resource management from a technical function to a strategic asset.

AI Workload Orchestration

AI Workload Orchestration seamlessly integrates with GPU Fractioning, distributing computing power across diverse workloads efficiently. This scheduler aligns resource distribution with business rules and priorities, ensuring technical and strategic alignment.

GPU Fractioning

GPU Fractioning addresses the efficient utilization of compute resources, particularly GPUs, by dynamically adapting to fluctuating workload demands. This is achieved through innovative features like GPU Fractioning, which subdivides physical GPUs into smaller, logical units, ensuring that workloads receive necessary computational power without the need for additional hardware.

The Leader In AI Infrastructure Management

Key Results

10×
GPU Availability

20×
Workloads Running

5×
GPU Utilization

0
Manual Resource Intervention

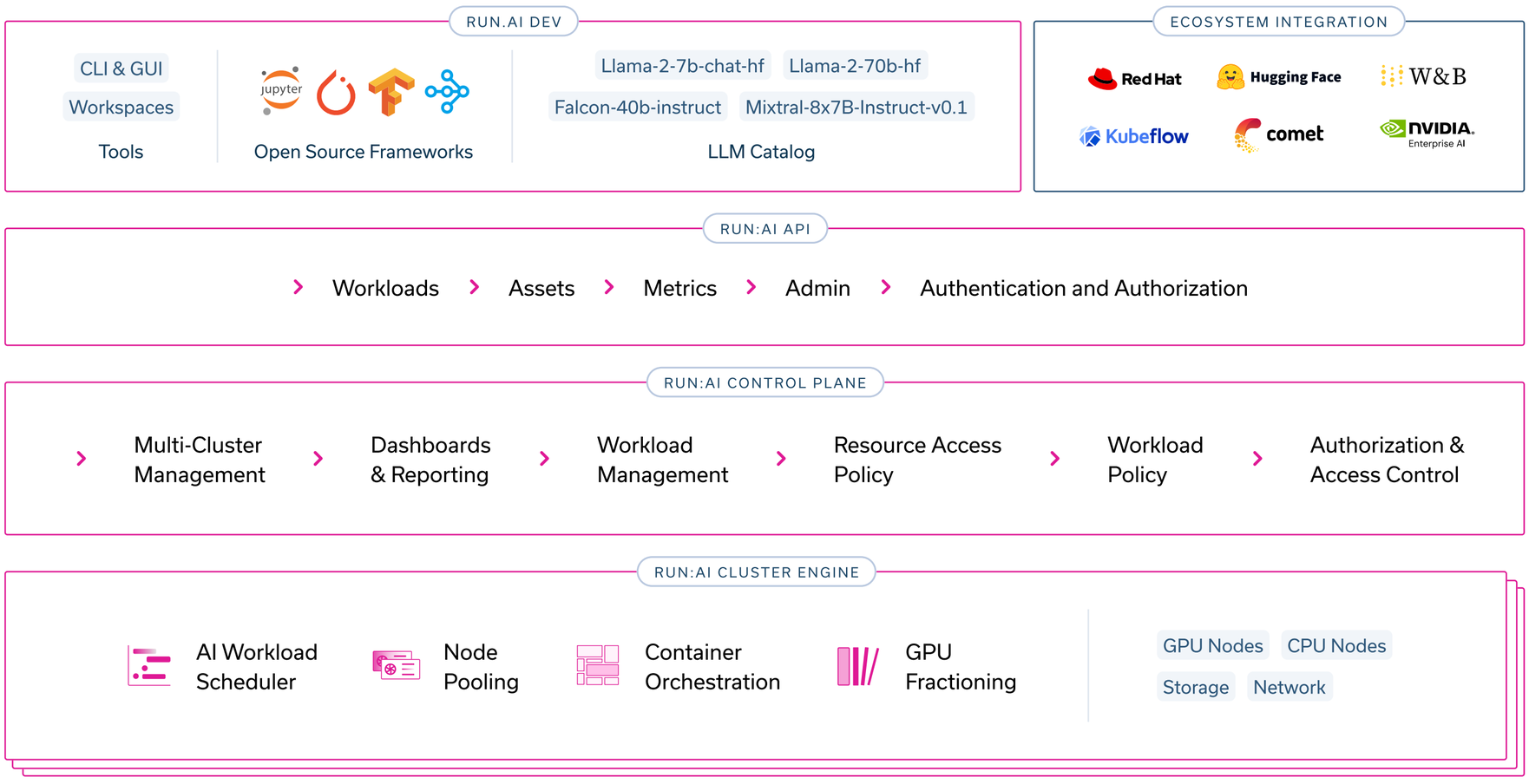
Leading U.S. bank using Run:ai platform

About Run:ai

Our mission is to significantly impact the AI revolution, pioneering positive change across industries for humanity. We provide the software layer implicit in empowering AI infrastructure and innovation. We enable the breaking of barriers and give people the tools to succeed both individually and collectively, revolutionizing the world with AI.

The Run:ai Platform

Run:ai offers the leading AI infrastructure management platform revolutionizing the way enterprises manage and optimize their AI and machine learning operations. The platform is specifically designed to address the unique challenges of AI infrastructure, enhancing efficiency, scalability, and flexibility.



Run:ai Dev - Comprehensive AI Lifecycle Support

Run:ai Dev empowers data scientists and AI researchers by providing a holistic, end-to-end platform that supports every phase of the AI development lifecycle. From initial model conceptualization to deployment, Run:ai Dev offers interactive, flexible environments with on-demand access to a wide range of tools and frameworks. This component ensures a seamless, efficient journey through the AI lifecycle, enabling innovation and accelerating the path from idea to implementation

Run:ai Control Plane - Resource Pooling and Policy Engine

Run:ai Control Plane is the central hub for managing and optimizing your AI resources. It pools compute resources across multiple environments - be it cloud, on-premises, or hybrid setups - into a single, manageable entity. Beyond just aggregation, the Control Plane integrates sophisticated business rules and governance strategies, ensuring that resource allocation is not only efficient but also aligned with your enterprise's strategic objectives. This component transforms resource management from a basic operational task into a strategic advantage, driving efficiency and agility in your AI operations

Run:ai Cluster Engine - AI Workload Orchestration and GPU Fractioning

The Run:ai Cluster Engine is at the forefront of intelligent workload management and resource optimization. It dynamically adjusts computing resources to meet the demands of diverse AI workloads, ensuring optimal utilization of GPUs. The Cluster Engine integrates with Run:ai's advanced scheduling capabilities to distribute computing power effectively, in accordance with the established business rules and priorities. This ensures that every AI task, from development to inference, receives the right amount of compute power, maximally utilizing the available resources and boosting overall throughput and efficiency

Run:ai Open Architecture - API and Ecosystem

The Open Architecture pillar of the Run:ai platform combines the strengths of API flexibility and a rich ecosystem to offer a versatile, integrative infrastructure. This feature empowers users with robust APIs, enabling seamless integration with a wide array of external tools, systems, and cloud environments. It ensures that Run:ai can effortlessly connect with and leverage the best of what the evolving AI technology landscape has to offer. Additionally, the Open Architecture signifies Run:ai's commitment to fostering a collaborative ecosystem, involving partnerships with leading cloud providers, AI technology vendors, and support for various open-source frameworks. This approach not only enhances the platform's functionality and adaptability but also keeps it at the forefront of AI innovation, providing users with a comprehensive, future-proof solution for their AI infrastructure needs