DATASHEET

ZEDEDA Edge Inference Services

Intelligence Where It Matters Most

Rapidly create, test, and deploy autonomous edge agents with any AI model on any edge hardware platform.

ZEDEDA Edge Inference Services

It combines a breakthrough approach to building and orchestrating Edge AI, comprising agents, models, applications, and infrastructure. Using it, Enterprise AI teams can rapidly create, test, and deploy autonomous edge agents with any AI model on any edge hardware platform, then operate them with the same edge orchestration already trusted by the world's largest enterprises.

Autonomous Agent Use Cases

Edge AI agents don't just analyze data — they act on it. Consider these examples, all of which take action based on visual AI:

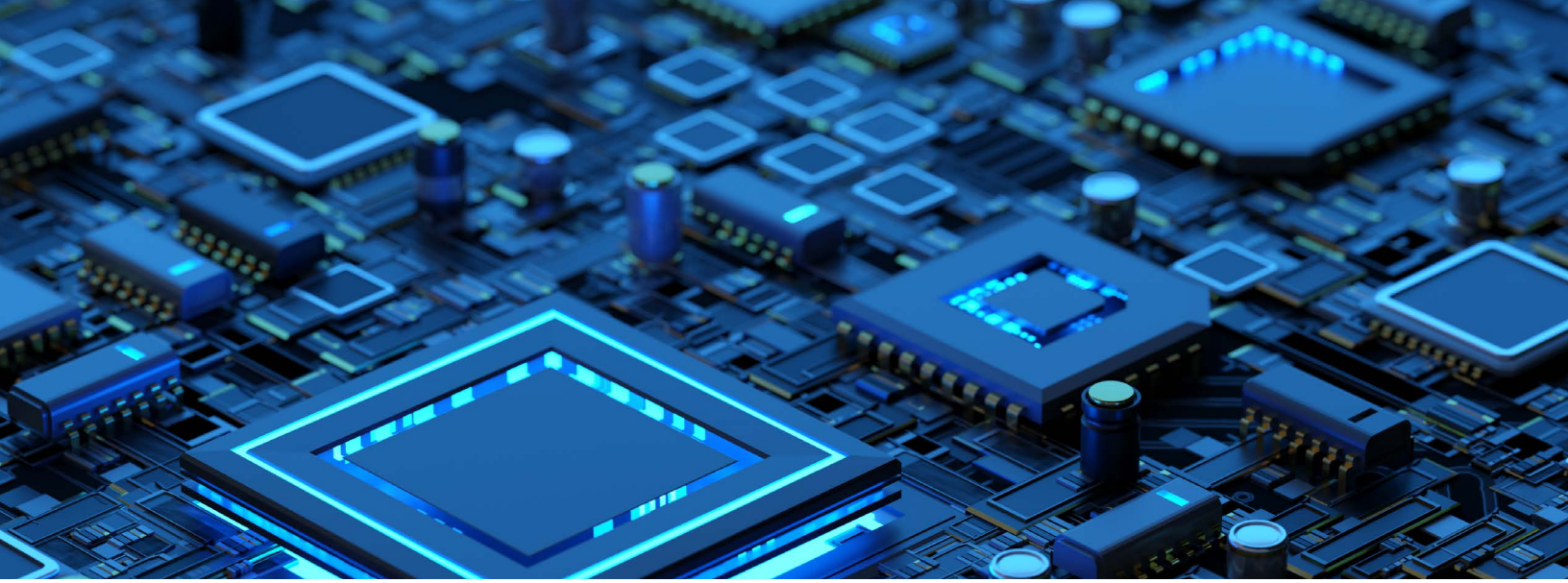
Manufacturing
Flag defective parts, then direct a robotic arm to pull them off the line.

Retail
When a store queue hits three customers, automatically alert the nearest available employee to open another register.

Energy
When pressure drives an oil flare beyond safe limits, immediately shut down the pump to protect workers and equipment.

Logistics
When a container arrives damaged, read its tracking ID and automatically route an inspection request to the right shipping company.

ZEDEDA



Pre-Validated Agentic Solutions

Use pre-validated, reusable, version-controlled Agentic Solutions. These are packages that bundle everything you need to run an AI workload on the edge: model artifacts, Helm charts, configuration values, deployment metadata, and hardware.

Share, version, and promote Agentic Solutions across teams and environments without manually re-creating them. Each one enforces consistent configuration for resource limits, inference server selection, and platform-specific settings, making it easier to scale and manage Edge AI.

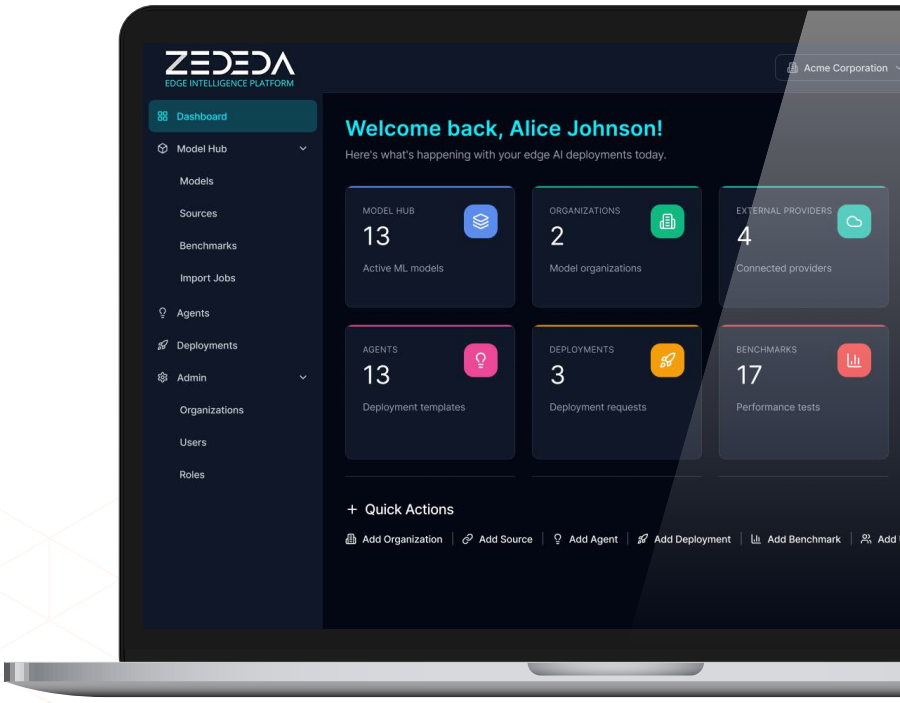
Custom Agentic Solutions

Build Agentic Solutions tailored to your operations, engage with them to refine and test behavior, then monitor their health and lifecycle.

AGENTIC AI APPS

AGENT OBSERVABILITY

AGENT LIFECYCLE



Streamline Edge Inference

Automate your entire Edge MLOps pipeline, from model import to production monitoring.

Capability	What It Does	Business Value
Access	Import models for your team to use	Leverage existing models for use at the edge, reducing engineering effort
Version	Track model version and lineage	Ensure traceability and reproducibility across your edge MLOps lifecycle
Validate	Benchmark model performance on actual edge hardware, including NVIDIA and Intel silicon	Know your device sizing, costs, and performance before deployment
Package	Automatically bundle inference based on hardware architecture, including OpenVINO, NVIDIA Triton, vLLM, and Ollama	Reduce engineering effort, promote reuse, and enforce best-practice configurations
Govern	Enforce GitOps workflows with approval gates for model governance and deployment	Ensure automated, reliable roll-outs across edge fleets
Optimize	Accelerate model performance on-device	Train models on high-end cloud silicon, then run efficiently on affordable edge silicon
Secure	On-device encryption, remote attestation, and hardware-based root of trust	Protect sensitive PII, as well as valuable code and model weights, even if a device is stolen
Operate	Monitor model performance across your edge fleet	Understand the status and performance of models and hardware to perform rapid troubleshooting of production edge systems

ZEDEDA

Accelerate Edge AI Deployments

Push your Edge AI projects beyond the pilot phase and into production – fast, safe, and secure.

	Before Edge Inference	With Edge Inference
Deployment Times	Weeks per model	Minutes per model
DevOps Team Involvement	Mandatory for every deployment	None; self-service
Infrastructure Setup	Manual, slow, error-prone	Automatically generated
Performance Benchmarking	None; hope model performs well	Models benchmarked on actual hardware
Platform Dependencies	Platform-specific pipelines for x86, ARM, NVIDIA	One pipeline for all platforms
Governance	None; no control over model versioning or branching	GitOps-based versioning of all model versions and branches
Compliance Risk	Sensitive data sent to cloud	Sensitive data stays on device, encrypted

Architected for Edge AI

Built for the requirements of Edge AI – not retrofitted from a cloud platform.

Requirement	Description
Model Registry	Pull models from wherever they live: NVIDIA NGC, Hugging Face, AWS S3 and SageMaker, Azure ML and Blob Storage, MLflow, or your local file system.
Lifecycle Management	Track model lineage, versions, deployment stages, and tags. Track model performance trends, inference frameworks, and dependencies. Know exactly which devices are running which model — and which ones need an update.
Organizations	Target deployments by location type — retail stores, production lines, warehouses, and more.
Jupyter Notebook Integration	Manage repositories, models, agents, and deployments programmatically via Jupyter Notebooks and ZEDEDA Edge Inference API.

Access Control	Role-based access control across every layer: models, repositories, organizations, deployments, devices, benchmarks, and agents.
A/B Testing	Test models in production with actual sensor data and push independent upgrades automatically without disrupting operations.
Real Hardware Benchmarking	Compare model performance on real edge devices using latency, throughput, resource utilization, power consumption, temperature, and reliability metrics.
Comprehensive Performance Monitoring	Monitor every edge cluster across all locations via map or list view. Aggregate metrics, including latency, throughput, and utilization. Drill into detailed telemetry and logs when you need to go deeper.

jupyter Hugging Face aws Microsoft Azure mlflow kubernetes

To learn more about how we can help you deliver intelligence to the edge, contact us today.

ZEDEDA

CONTACT@ZEDEDA.COM

About ZEDEDA

ZEDEDA unlocks the value of AI where it matters most, enabling enterprises to create, secure and operate edge AI at scale. ZEDEDA's Edge Intelligence products and solutions are used by global distributed enterprises to rapidly realize and deploy autonomous intelligence wherever they operate, turning real-time data into real and tangible business outcomes and decisions. Trusted by the world's largest organizations, ZEDEDA is backed by world-class investors, with teams in the United States, Germany, India, and the United Arab Emirates. For more information, visit www.ZEDEDA.ai.