Cloudian and AWS Bring High-Performance AI Inferencing to the Edge with HyperScale AI Data Platform on AWS Local Zones

Pay-as-you-go infrastructure enables enterprises to run RAG applications at the edge with single-digit millisecond latency

SAN MATEO, CA, UNITED STATES, October 28, 2025 /EINPresswire.com/ -- Cloudian, the leader in enterprise object storage solutions, today announced Cloudian HyperScale® AI Data Platform on AWS Local Zones, delivering breakthrough edge AI inferencing capabilities with rapid deployment and pay-as-you-go infrastructure pricing. This collaboration brings NVIDIA GPU-accelerated AI workloads to the edge, enabling enterprises to process sensitive data locally while maintaining single-digit millisecond latency to end users.

AWS Local Zones place compute, storage, and select AWS services closer to large population centers and industry hubs. By deploying HyperScale AI Data Platform on AWS Local Zones, enterprises gain immediate access to high-performance AI infrastructure without building data centers, while maintaining the operational simplicity and pay-as-you-go economics of AWS.

The initial use case for HyperScale AI Data Platform on AWS Local Zones focuses on Enterprise Document RAG (Retrieval Augmented Generation)—making decades of institutional knowledge instantly accessible through a familiar chatbot interface. Organizations can deploy AI agents that understand and reason over their complete repository of documents, manuals, reports, and multimedia content stored in S3-compatible formats.

"Organizations no longer need to choose between performance, data sovereignty, and cost efficiency," said Neil Stobart, CTO at Cloudian. "By combining Cloudian's high-performance S3-compatible storage with AWS's GPU-based edge infrastructure, we're enabling enterprises to run sophisticated RAG applications within milliseconds of their end users, with zero upfront cost. This dramatically accelerates AI adoption for organizations that previously couldn't justify the

infrastructure investment."

The benefits are transformative: Customer service teams access product documentation instantly for accurate responses, field technicians retrieve repair procedures in real-time, and employees find answers without navigating complex file systems. The result is enhanced customer experience and increased employee productivity.

The HyperScale AI Data Platform deployment on AWS Local Zones includes GPU-powered cloud servers purpose-built for demanding AI/ML workloads, featuring up to eight NVIDIA Hopper GPUs with 640 GB of GPU memory, 3rd Gen AMD EPYC processors, and 3,200 Gbps of Elastic Fabric Adapter (EFA) networking for massive scale-out performance. This configuration enables real-time inferencing with sub-10 millisecond response times while processing sensitive data locally.

"Enterprise AI success at the edge requires bringing high-performance compute and storage together where data needs to be processed," said Justin Boitano, vice president, Enterprise AI Products at NVIDIA. "Cloudian's HyperScale AI Data Platform integration with NVIDIA accelerated computing on AWS Local Zones enables organizations to build and scale intelligent RAG applications to power AI agents and reasoning workloads close to their data."

Traditional edge AI deployments require substantial capital investment in data centers, GPU servers, and networking equipment—often requiring 6-12 months from planning to production. AWS Local Zones eliminates these barriers with pay-as-you-go pricing, enabling organizations to launch production-ready AI infrastructure in hours using familiar AWS tools and APIs. AWS Local Zones are available in 35 metropolitan areas around the world, with GPU-accelerated instances for AI workloads available in select locations. For availability details, visit the AWS Local Zones features page.

AWS Local Zones extend AWS infrastructure closer to end-users in metropolitan areas, enabling ultra-low latency applications critical for government and healthcare. Public sector agencies use them for real-time citizen services and emergency response systems, while health sciences leverage them for medical imaging, telemedicine, and patient monitoring requiring immediate data processing and regulatory compliance.

Key Capabilities

 Edge AI Performance: Single-digit millisecond latency with up to eight NVIDIA H100 GPUs per instance for real-time inferencing.

 S3-compatible storage: Cloudian’s enterprise-proven native S3 architecture ensures full compatibility with S3-based applications and data.

 Integrated Vector Database: Built-in capabilities automatically ingest, embed, and index

multimodal content for immediate RAG deployment.

 Data Sovereignty: Process sensitive data locally while maintaining compliance with regional data residency requirements.

 Pay-as-you-go pricing: AWS pay-as-you-go pricing eliminates capital expenditure, enabling deployment at less cost.

 Rapid Deployment: Launch in hours instead of months, dramatically reducing time-to-value.

By eliminating infrastructure complexity and capital investment barriers while maintaining enterprise-grade performance and data sovereignty, Cloudian HyperScale AI Data Platform on AWS Local Zones enables organizations of all sizes to harness the transformative power of edge AI for their business-critical applications. For more information, visit www.cloudian.com/aws.

Jon Toor
Cloudian
email us here
