NEWS RELEASE

Penguin Solutions Selected by Deepgram to Enable Deployment of Optimized AI Inference Infrastructure for Enterprise Voice AI

2026-03-17

Strategic collaboration leverages Dell PowerEdge servers and NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs to deliver high-performance, low-latency voice experiences for mission-critical applications in healthcare and retail

FREMONT, Calif.--(BUSINESS WIRE)-- **Penguin Solutions**, Inc. (Nasdaq: **PENG**), the AI factory platform company, today announced a strategic collaboration with Deepgram and Dell Technologies to architect and deploy a fully optimized, production-ready infrastructure aligned to Deepgram's demanding enterprise voice AI requirements. By leveraging its unique expertise in designing, building, deploying, and managing **AI infrastructure** with Dell PowerEdge servers and Dell PowerScale storage optimized for AI workloads, Penguin Solutions delivered an optimal solution to support and enhance Deepgram's innovative Speech-to-Text (STT), Text-to-Speech (TTS), and Voice Agent capabilities, while ensuring maximum reliability and performance.

Penguin Solutions strategic collaboration with Deepgram and Dell Technologies on a fully-optimized, production-ready infrastructure delivers an optimal solution to support and enhance Deepgram's innovative Speech-to-Text (STT), Text-to-Speech (TTS), and Voice Agent capabilities. The solution aligns to Deepgram's demanding enterprise voice AI requirements while ensuring maximum reliability and performance.

As enterprise adoption of generative AI accelerates, organizations must adhere to stricter service level agreements (SLAs), which require infrastructure that can ensure low latency and high concurrent usage. This Penguin-led deployment addresses these challenges by combining Deepgram's innovative voice AI models with a purpose-built architectural design, a highly efficient deployment, and ongoing performance optimization.

"Modern AI workloads demand infrastructure that performs consistently and scales predictably under heavy loads,

particularly for real-time inference applications like voice agents," said Joe Castillo, vice president of sales at Penguin Solutions. "By partnering with Deepgram and utilizing proven Dell AI infrastructure, Penguin Solutions is delivering a validated, scalable, end-to-end architecture. Our comprehensive framework equips Deepgram with the optimized infrastructure needed to reliably and accurately deliver complex voice AI capabilities in healthcare, retail, and other industries."

Drawing on its extensive experience with HPC and AI infrastructure, Penguin Solutions ensures that the underlying infrastructure meets the specific demands of Deepgram's neural networks. The architecture also incorporates Dell PowerScale storage and Dell PowerEdge XE7745 servers with **NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs**, which provide efficient inferencing that enables data-intensive voice applications to operate seamlessly in real-time environments.

"Deepgram is focused on delivering voice AI capabilities that meet the demanding performance, scalability, and reliability requirements of enterprise environments - something only Deepgram brings to the market today," said Abe Pursell, vice president of partnerships and business development at Deepgram. "The infrastructure behind our platform has to be equally robust to support that level of innovation. Penguin Solutions demonstrated a deep understanding of our technical requirements, translating them into a sophisticated infrastructure environment that meets and exceeds expectations. This enables us to continue delivering the enterprise-class capabilities our customers rely on."

"AI-driven voice applications are transforming how organizations engage with customers and patients, but success depends on a resilient, high-performance infrastructure foundation," said David Noy, vice president, unstructured data solutions product management at Dell Technologies. "Our collaboration with Penguin Solutions demonstrates how AI-optimized Dell PowerScale storage and Dell PowerEdge servers with NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs can accelerate enterprise AI adoption at scale. Together, we're enabling Deepgram to deliver secure, low-latency voice AI experiences that power mission-critical innovation across healthcare and retail."

The Deepgram-Penguin Solutions-Dell collaboration comprises a comprehensive approach for enterprises looking to modernize their customer and employee experiences. With Deepgram's API-driven voice capabilities, Penguin Solutions' AI services, and Dell's powerful AI infrastructure, organizations can achieve highly accurate, real-time transcription and speech synthesis—all while maintaining strict data governance and control.

For those attending NVIDIA GTC AI Conference and Expo March 16-19, 2026, in San Jose, CA, learn more about this innovative collaboration at Dell's Booth #721 on March 17 at 3:30 p.m. for the session "Powering Enterprise Voice AI: Deepgram's Agentic Solution" presented by Penguin, Deepgram and Dell. Attendees can also stop by Penguin Solutions' booth #1031 to speak with an AI factory platform expert.

About Penguin Solutions

The most transformative technological advancements are often the hardest to deploy and optimize. Penguin Solutions, the AI factory platform company, has the innovative technologies, skills, experience, and partnerships needed to turn your AI ambitions into reality.

In addition to our AI capabilities, Penguin Solutions offers memory and LED solutions serving a wide range of high-performance and specialized applications.

For more information, visit **https://www.penguinsolutions.com**.

PR Contact

Maureen O'Leary
Penguin Solutions
Corporate Communications
1-602-330-6846
pr@penguinsolutions.com

Source: Penguin Solutions, Inc.