SambaNova SYSTEMS

USE CASE

SambaNova Cyber-Attack Response Solution

World record performance and accuracy for secure generative AI

Challenge

The US Government needs to rapidly classify threats, develop responses, and deploy them before additional damage can occur, with challenges in automating security processes such as patch management.

Objective

To leverage SambaNova's generative AI capabilities to automate response protocols, enabling quick isolation of affected systems and mitigation of damage. AI-driven decision-making will be utilized to determine the best course of action for each specific threat, reducing response times and minimizing impact. SambaNova's powerful AI platform can handle the computational demands of real-time attack response, ensuring efficient and effective mitigation of cyber threats.

Distinguishing Factors

- Domain trained expert models
- Composable pipelines
- World record performance and accuracy
- Data security & privacy
- Model/data ownership
- Full stack solution
- On-site/cloud services

Justification

The SambaNova generative AI platform can automate response protocols quickly, as demonstrated by its "world record 1000 Tokens/sec inference at 1T parameter scale." This speed is essential for isolating affected systems and mitigating damage rapidly in the event of a cyber-attack.

Solution

SambaNova Suite, combined with the Samba-1 Composition of Experts (CoE) model, delivers the highest combination of performance and accuracy for generative AI. A complete hardware/software platform, SambaNova Suite can be deployed as a fully configured rack-level solution either on-premises, including air-gapped environments, or as a cloud-based solution. The Samba-1 model provides state-of-the-art accuracy across a wide range of use cases.

Key Features

Performance, Accuracy, and Efficiency
Only SambaNova delivers performance of over 1000 tokens/s at full precision, on as few as 16 sockets. This is an unrivaled combination of performance and accuracy with a small footprint, dramatically reducing power consumption.

Total AI Stack Control
With options for on-premises and air-gapped deployments, SambaNova ensures complete data privacy and security, enabling the use of sensitive data for model training without external exposure.

Model Ownership
Once a model is fine-tuned, it becomes the property of the customer in perpetuity, eliminating model ownership concerns.

Configurable Access and Permissions
The Samba-1 model enables flexible access controls, allowing precise management of model permissions to align with organizational roles, enhancing security and ensuring compliance.

SambaNova Suite delivers the most accurate, integrated full stack generative AI platform, optimized for enterprise and government organizations. Delivered through SambaNova's integrated AI platform, which can be deployed on-premises or through the cloud, SambaNova Suite seamlessly integrates into existing business processes to deliver transformative capabilities. With the capacity to be further refined using customer data, these models can deliver unparalleled accuracy while providing enterprise grade security and data governance.

Learn more