CORESTACK®

CoreStack Framework

Intelligent Governance for Generative AI at Scale

Driving Visibility, Control, and Accountability Across Multi-Cloud AI Workloads

By Perumal Balakrishnan, Senior VP – CoreStack

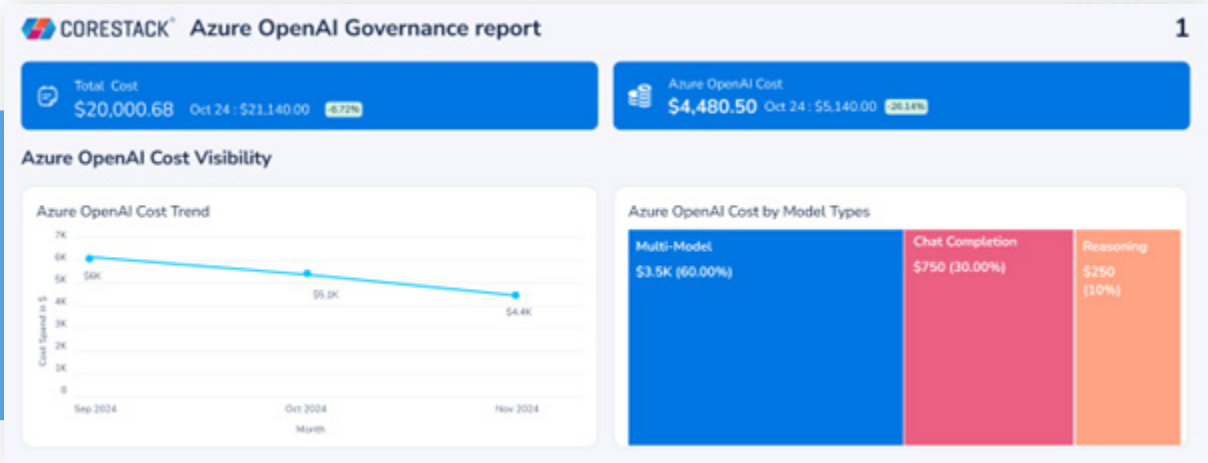
Executive Summary

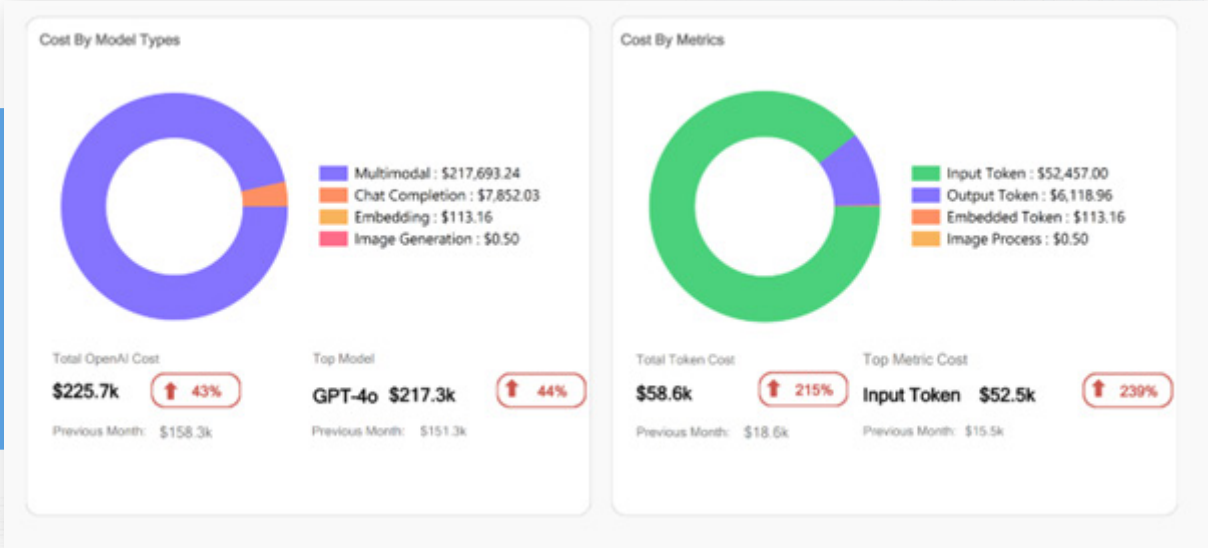
The rapid evolution of generative AI presents a compelling opportunity for businesses to enhance productivity, automate tasks, and unlock new avenues for innovation. However, deploying generative AI solutions at scale requires a robust governance framework to address critical challenges such as security vulnerabilities, cost overruns, compliance risks, and ethical concerns.

A robust governance framework is essential for successfully implementing generative AI platforms. This framework should prioritize cost management, data security and operational efficiency. The CoreStack Governance framework for GenAI services addresses the key challenges of implementing effective AI governance.

Visibility

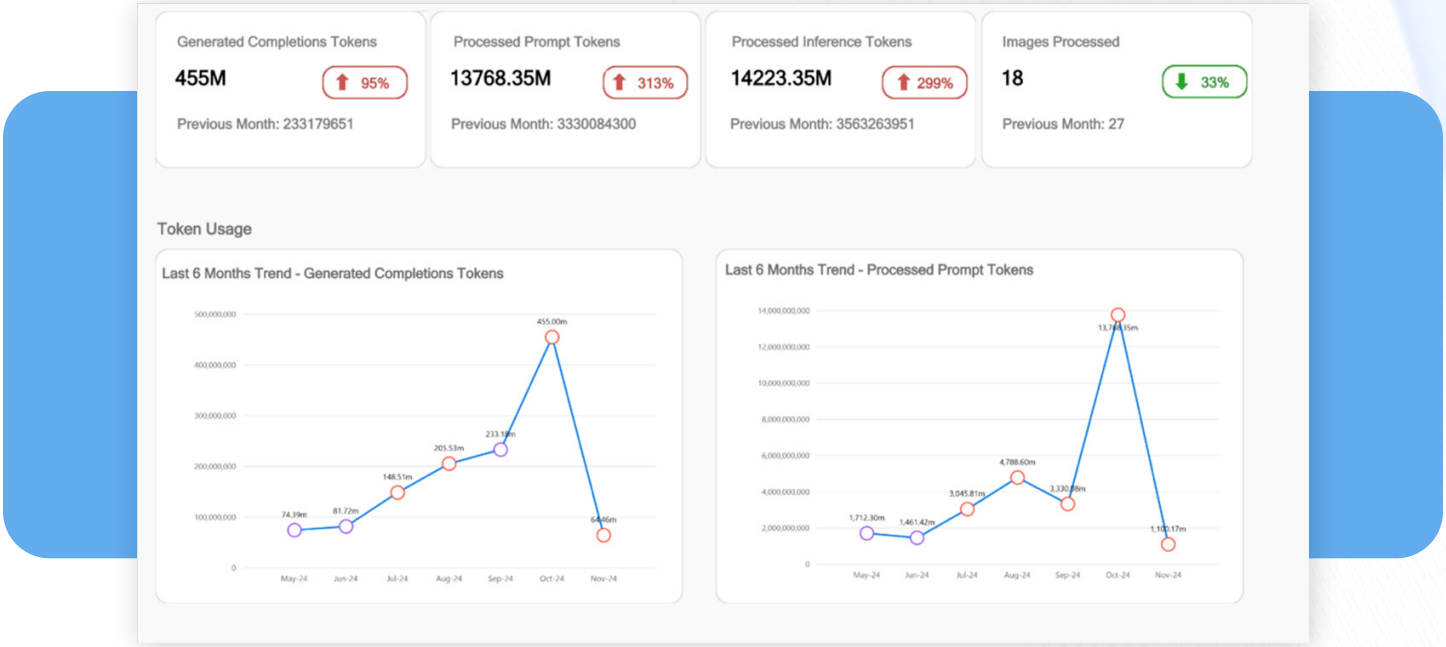
CoreStack Framework enables a comprehensive visibility to monitor and allocate costs by business unit, application, and user, including enabling detailed dashboards to track Input Token Count, Output token count, Image Processed count, Processed Inference count and Model training hours grouped based on Application and Business group.





Model Usage Metrics

The platform tracks tokens processed (inputs and outputs), model invocation counts, training hours, and specialized metrics like image generations or characters processed in speech models.



Token Cost:

Token costs for model inference and fine-tuning across all Gen AI cloud services.

Top Cost Drivers

Top consuming models, datasets, and jobs visibility.

Cloud Cost Comparison

Cross-cloud comparisons enable organizations to compare AWS Bedrock vs. Azure OpenAI costs for similar use cases, helping inform cloud strategy and vendor management.

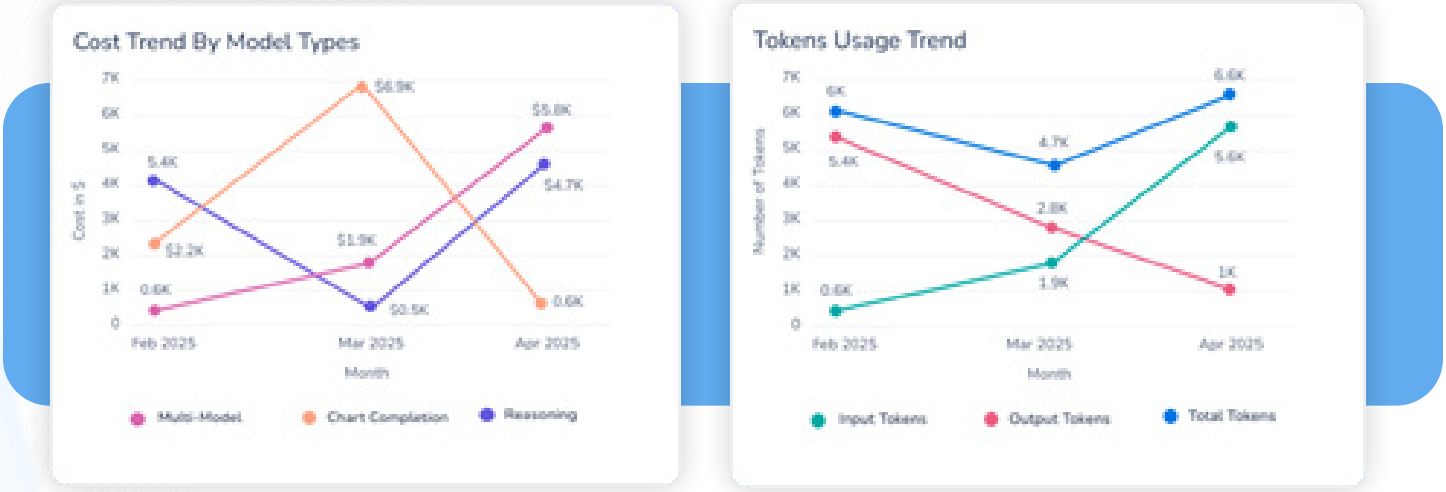
Track Cost Trends by Models and Token Usage

Visualize cost trends over time, segmented by individual models and overall token consumption patterns. This allows for proactive identification of potential cost overruns and optimization opportunities.

- Model-Specific Costs
- Input Token Count
- Output Token Count
- Total Token Count

Model-wise Input and Output Token Count

Detailed Token Consumption per Model: Gain granular visibility into the number of input and output tokens processed by each model. This allows for precise cost attribution and understanding of model usage patterns.



Custom Dataset Training Hours

Monitor the time spent specifically on training custom datasets for models. This provides insights into the efficiency and cost implications of custom model development.

Image Processing

Total Image Processed Count with Resolution Type: For image-based models, measure the total number of images processed and further categorize them by resolution. This allows for understanding the relationship between image complexity and processing costs

High Utilization Models

Identify high Utilization Models, Regions, and Subscriptions: Gain visibility into resources experiencing high utilization across models, regions, and subscriptions. This allows for proactive resource allocation and potential cost optimization by leveraging reserved instances or other cost-effective strategies.

Assessment

Conducting a GenAI Governance Assessment is vital for organizations seeking to implement best practices and automate rules tailored to their unique governance needs. The selection of a Generative AI foundational model hinges on several critical factors, including the specific task at hand, desired accuracy, expected performance, the supported number of input/output tokens, and the extent of multi-language support required. Inference costs significantly impact the overall expenses associated with Generative AI products. Thus, it's crucial to architect solutions efficiently to minimize token counts while retaining response accuracy. Utilizing the most suitable foundational models can enhance value and cost-effectiveness in AI deployments.

The CoreStack framework enables cost optimization opportunities tailored to the organization needs, covering the scenarios and best practices outlined below.

Below are several scenarios to guide the selection of well-architected Generative AI models based on application needs:

Language Generation and Complex Query Handling

For applications needing advanced context and nuanced reply generation, using models like GPT-4 is advisable. For instance, a customer service chatbot that handles complex issues will benefit from GPT-4's extensive processing capabilities.

Basic Task Automation

For simpler tasks like data entry or straightforward queries, models like Babbage-002 may suffice. They ensure cost-effectiveness while delivering satisfactory performance.

Fine-tuning for Specialized Use Cases

Organizations requiring customized performance can explore fine-tuning options. For instance, a company developing a unique customer support model can gain better results using targeted training with models like GPT-3.5-Turbo (4K) or Davinci-002 for more complex needs.

Multi-Language Support Applications

When developing international solutions, choosing a model offering extensive multi-language support can impact performance and cost. Applications requiring diverse language comprehension should look to models optimized for multi-language processing, potentially favouring GPT-4 for its higher performance in complex language scenarios.

Speech Models

Speech models like TTS and TTS HD incur costs based on usage hours and character volume. To manage costs, minimize idle time and keep text inputs concise. Use TTS for general tasks and TTS HD for high-quality speech needs.

Embedding Models

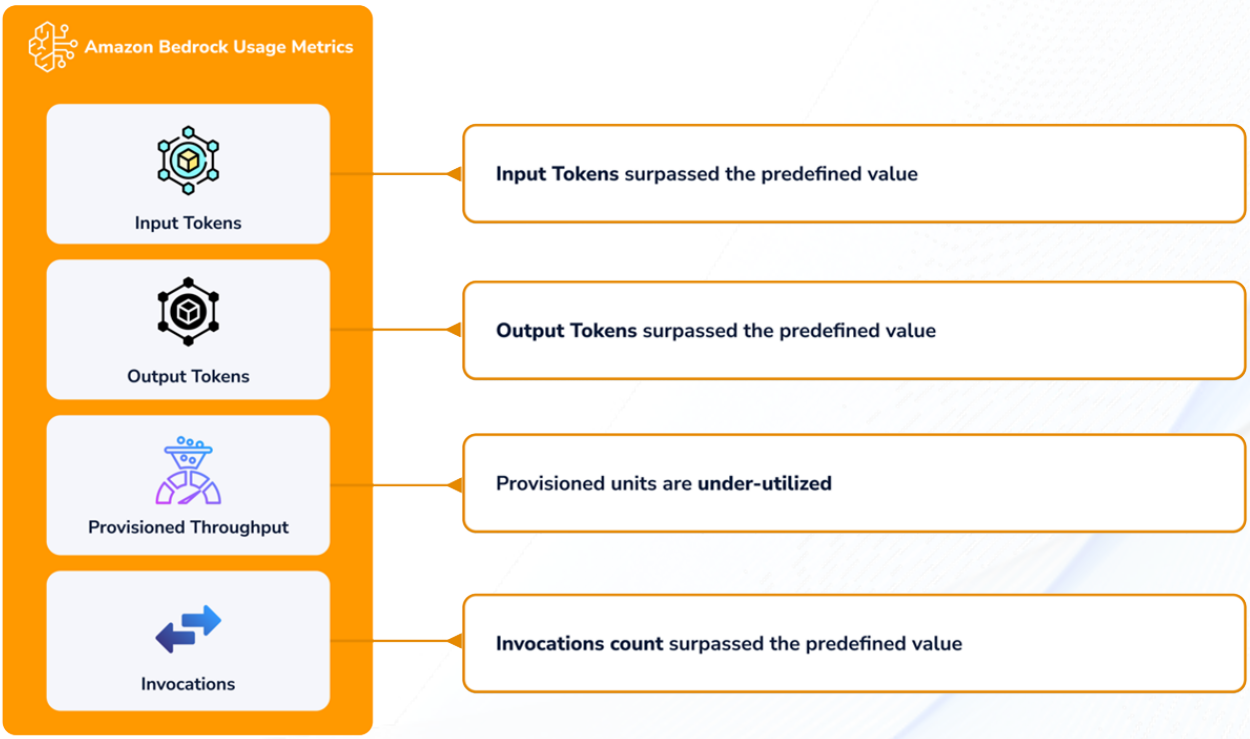
Embedding models include the Ada and Text-Embedding-3 series. Efficient data representation and trimming unnecessary data can reduce expenses. Ada is ideal for basic tasks, while Text-Embedding-3-Large and Text-Embedding-3-Small are suited for more complex or large-scale tasks.

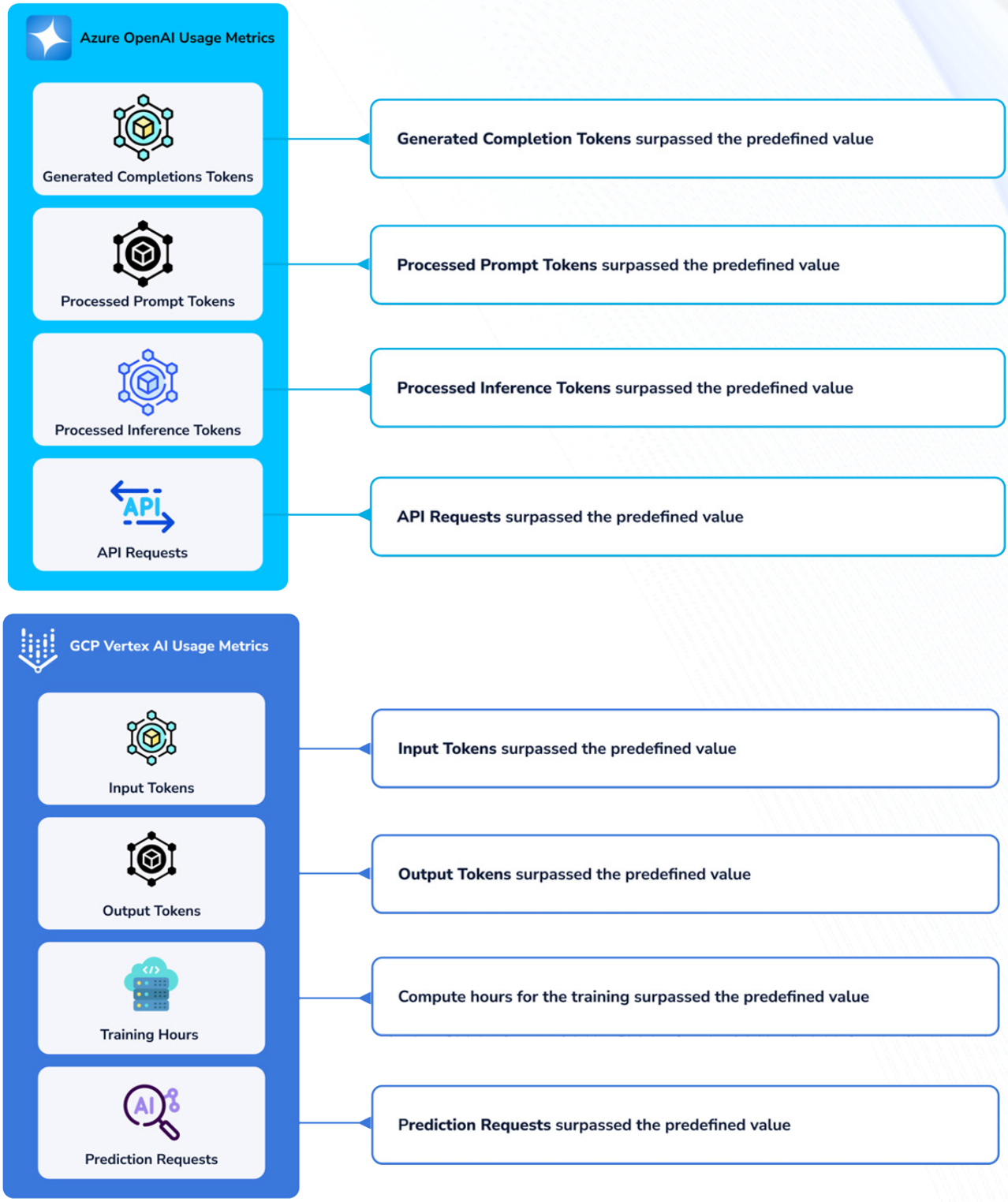
Fine-tuning Models

These models, such as Babbage-002 and GPT-3.5-Turbo, involve hosting, training hours, and token usage charges. Optimize training using relevant data, limit training duration, and manage token efficiency. For simpler fine-tuning, use Babbage-002 or GPT-3.5-Turbo (4K); for more complex tasks, opt for Davinci-002 or GPT-3.5-Turbo (16K).

Guardrails

Organizations need more than just monitoring and advice to govern AI at scale effectively; they require enforceable guidelines that ensure AI usage stays within safe and approved limits. CoreStack provides a framework for Guardrails Notification and Policy Enforcement that initiates with the deployment of AI services and their lifecycle management. This framework is designed to accommodate the diverse needs of the organization's various applications, environments, and users. It acts as an automated safety net for AI resource provisioning and lifecycle management. A robust guardrails notification framework is essential for proactively managing costs, enforcing usage policies, and mitigating potential risks associated with generative AI.





Anomaly

Configure anomaly detection mechanisms to identify and notify users about unusual patterns in model cost, token consumption, and other relevant metrics. This could include unexpected spikes in usage for specific Bedrock models, deployments, or even across your entire organization.

Budget

Establish budget thresholds for your organization, specific business units, and individual models. Configure alerts to notify stakeholders when usage approaches or exceeds these predefined limits, allowing for proactive cost management.

Quota

Define and implement quotas for various aspects of usage, such as the number of requests to specific models, the volume of text processed, or the total cost incurred. Configure alerts to notify relevant personnel when these quotas are nearing exhaustion.

API Rate Limiting

Implement API rate limits for the endpoints to prevent abuse, ensure fair usage across different users or applications, and protect against unexpected spikes in demand.

Access Control Violations

Configure Identity and Access Management (IAM) policies and alerts to monitor and prevent unauthorized access to the models, deployments, and associated training data resources.

Performance Degradation

Set up alerts to notify significant increases in inference latency or other performance degradations.

Resource Utilization

Track the utilization of models and deployments. Configure alerts for high resource consumption or potential bottlenecks, ensuring optimal resource allocation and preventing performance issues.

AI-Powered NextGen Cloud Governance and Security

Our NextGen Cloud Governance portfolio transforms organizations from a reactive posture to a proactive one, enabling predictable increases in top-line revenues and bottom-line efficiencies. To get started or to learn more, visit **www.corestack.io** or reach out to us at **sales@corestack.io.**

Industry Recognition

Gartner

Magic Quadrant: Cloud Management Platforms

FORRESTER®

Forrester Wave: Cloud Cost Management and Optimization

IDC

Innovator in Multicloud Management

FROST & SULLIVAN

Frost Radar: Leader, Global Hybrid Cloud Management

Deloitte. Technology Fast500

Deloitte Technology Fast 500 Winner

GeekWire

GeekWire 200 Winner

Inc. 5000

Inc. 5000 Winner

Inc. Best Workplaces

Inc. Best Workplaces Winner

CORESTACK®

CoreStack is an AI-powered NextGen Cloud Governance platform that enables enterprises to embrace cloud with confidence, rapidly achieving continuous and autonomous cloud governance at scale. CoreStack helps 750+ global enterprises govern more than $2B in annual cloud consumption. The company is a Microsoft Azure (Legacy), Gold Partner, Amazon AWS Technology Partner with Cloud Operations Competency, Oracle Cloud Build Partner, and Google Cloud Build Partner.