

The background features a dark navy blue field with abstract, overlapping shapes in vibrant magenta and deep red. Two thin, light blue lines intersect diagonally across the upper right portion of the image. The text is positioned on the left side.

AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

SEC323

The AWS approach to secure generative AI

Jason Garman

(he/him)

Principal Security Solutions Architect
Amazon Web Services

JD Bean

(he/him)

Principal Security Architect
Amazon Web Services



Agenda

1.

AWS approach to secure & responsible AI

2.

Security in the AWS generative AI tech stack

3.

Integrating AWS security services into your generative AI application

A decorative graphic consisting of several interlocking gears in purple and blue, and a white hexagon, positioned behind the text.

Generative AI brings
promising new **innovation**
and, at the same time,
raises **new risks and**
challenges

Responsible AI dimensions

FAIRNESS

Considering impacts on different groups of stakeholders

EXPLAINABILITY

Understanding and evaluating system outputs

CONTROLLABILITY

Having mechanisms to monitor and steer AI system behavior

SAFETY

Preventing harmful system output and misuse

PRIVACY AND SECURITY

Appropriately obtaining, using, and protecting data and models

GOVERNANCE

Incorporating best practices into the AI supply chain, including providers and deployers

TRANSPARENCY

Enabling stakeholders to make informed choices about their engagement with an AI system

VERACITY AND ROBUSTNESS

Achieving correct system outputs, even with unexpected or adversarial inputs

Decomposing responsible AI



Safety

- System design to prevent harmful system output, encourage accurate and relevant responses, and align with goals for fairness, bias, and desired brand voice
- Risks can only be *managed* through application of services such as Amazon Bedrock Guardrails
- Current systems can only address this risk probabilistically, and customers will need to *calibrate the acceptable risk* before deployment

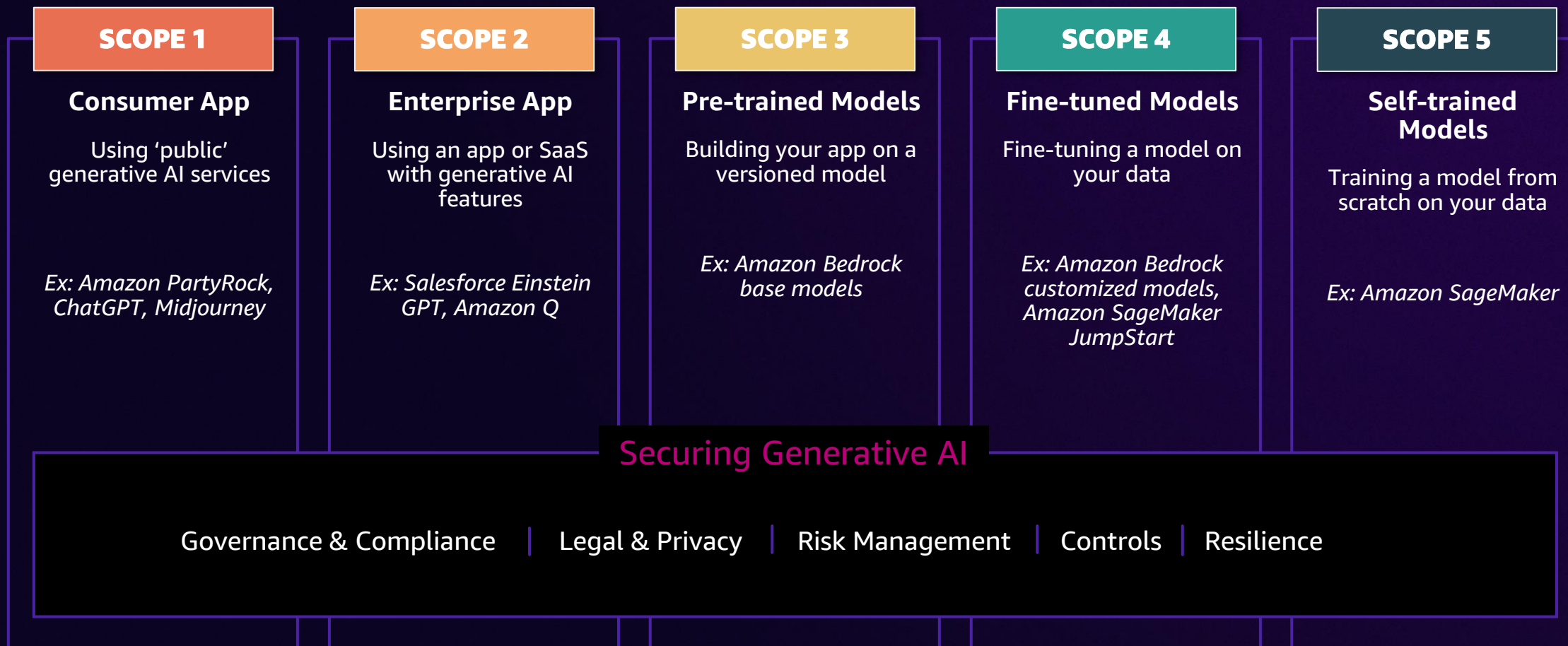


Security

- System design to prevent revealing sensitive/confidential information to unauthorized users
- Best practice is to use *traditional security controls* to filter all data before it enters the foundation model based on an end user identity
- Uses *deterministic, auditable, and explainable* controls
- *Any failure of the system* where users can be exposed to unauthorized content should be considered a blocker

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES



Security guidelines for generative AI

COMPLIANCE AND GOVERNANCE

The policies, procedures, and reporting needed to empower the business while minimizing risk

Create generative AI usage guidelines

Establish process for output validation

Develop monitoring and reporting processes

LEGAL AND PRIVACY

The specific regulatory, legal, and privacy requirements for using or creating generative AI solutions

Retain control of your data

Encrypt data in transit and at rest

Support regulatory standards

CONTROLS

The implementation of security controls that are used to mitigate risk

Human-in-the-loop

Explainability and audibility

Testing strategy

Identity and access management

RISK MANAGEMENT

Identification of potential threats to generative AI solutions and recommended mitigations

Threat modeling

Third-party risk assessments

Ownership of data, including prompts and responses

RESILIENCE

How to architect generative AI solutions to maintain availability and meet business SLAs

Data management strategy

Availability

High availability and disaster recovery strategy

Generative AI stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs



Amazon Q
Business



Amazon Q
Developer



Amazon Q in
QuickSight



Amazon Q in
Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs



Amazon Bedrock

Guardrails | Agents | Studio | Customization capabilities | Custom model import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



AWS Trainium



AWS Inferentia



Amazon SageMaker



Amazon EC2
UltraClusters



Elastic Fabric
Adapter



Amazon EC2
Capacity Blocks for ML



AWS
Nitro System



AWS
Neuron

Foundation layer: Infrastructure



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Generative AI stack: Infrastructure layer

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



AWS Trainium



AWS Inferentia



Amazon SageMaker



Amazon EC2
UltraClusters



Elastic Fabric
Adapter



Amazon EC2
Capacity Blocks for ML



AWS
Nitro System



AWS
Neuron

Securing AI infrastructure

ZERO ACCESS TO SENSITIVE AI DATA



[Read the blog](#)

1

Complete isolation of the AI data from the infrastructure operator

2

Ability for customers to isolate AI data from themselves

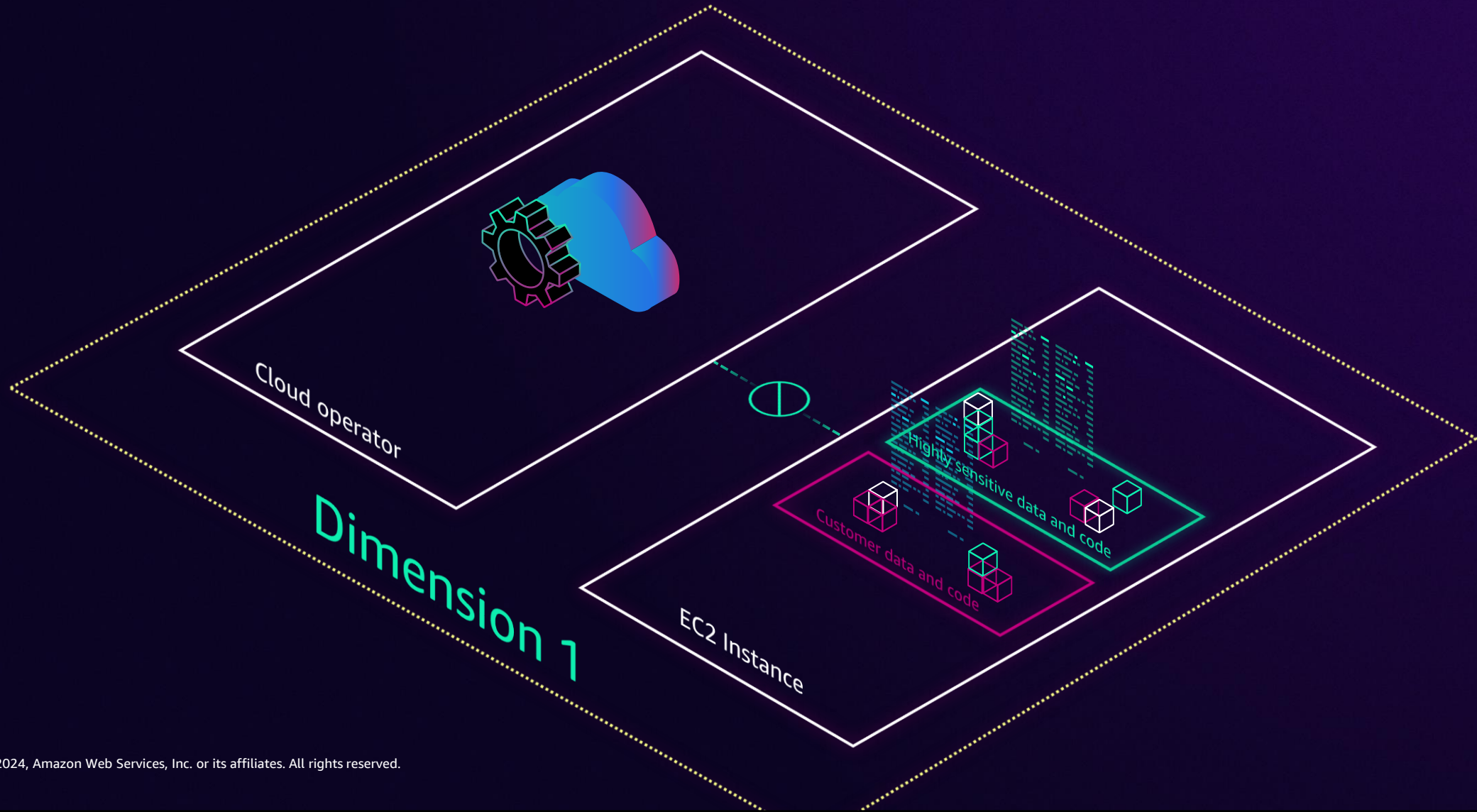
3

Protected infrastructure communication



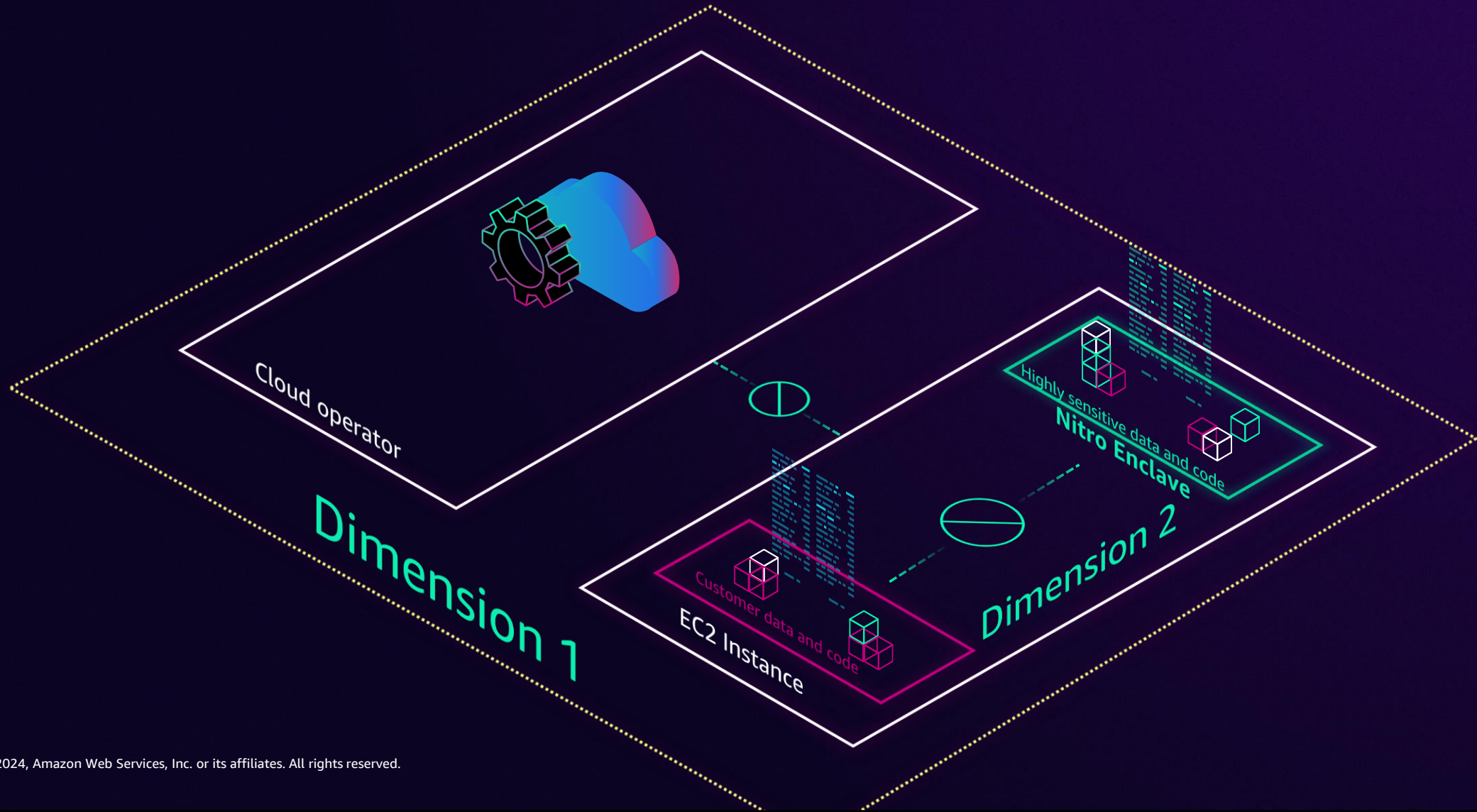
Confidential computing

PROTECTING CUSTOMER CODE AND SENSITIVE DATA IN USE



Confidential computing

PROTECTING CUSTOMER CODE AND SENSITIVE DATA IN USE

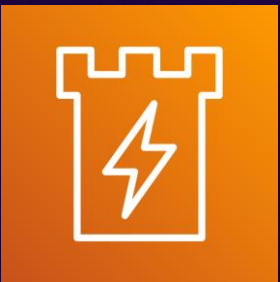


Confidential computing capabilities

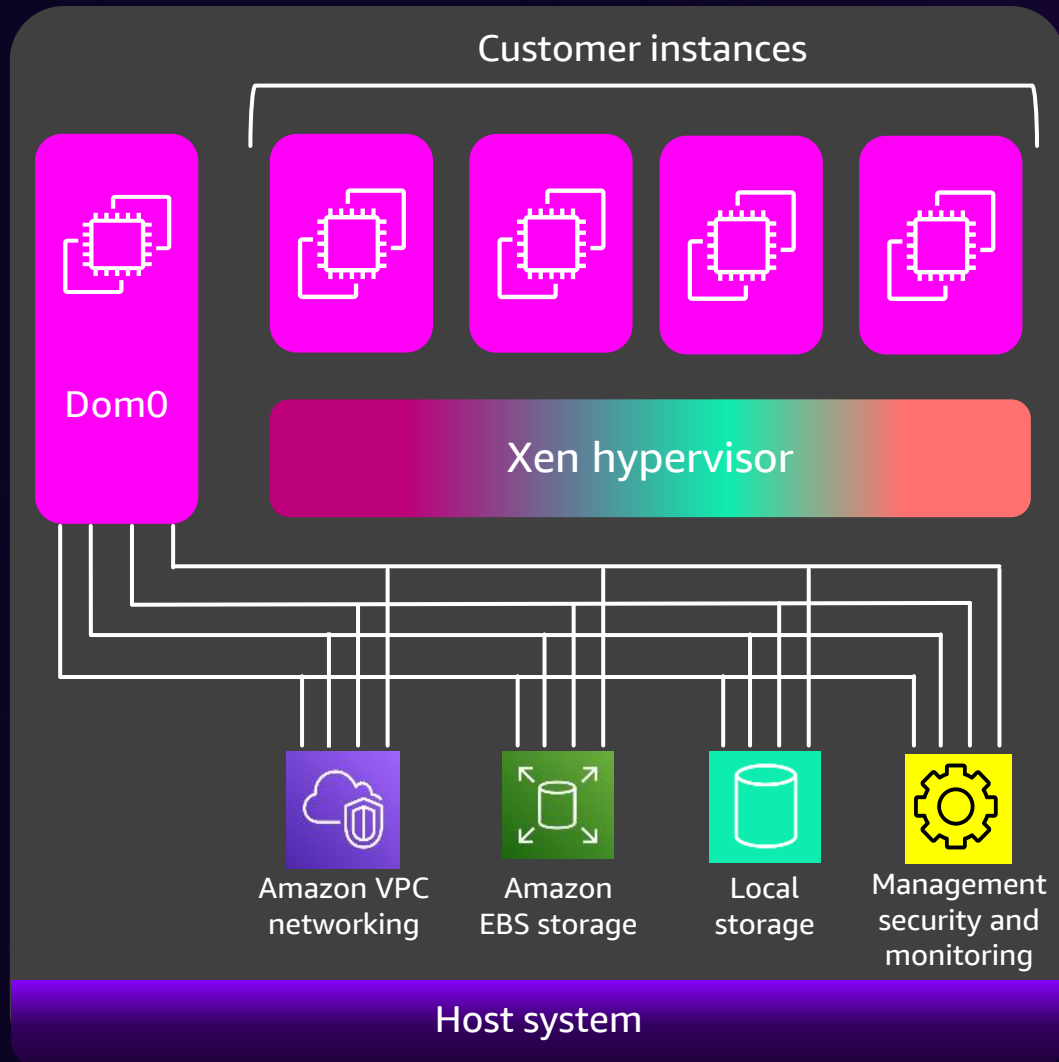
AWS Nitro System



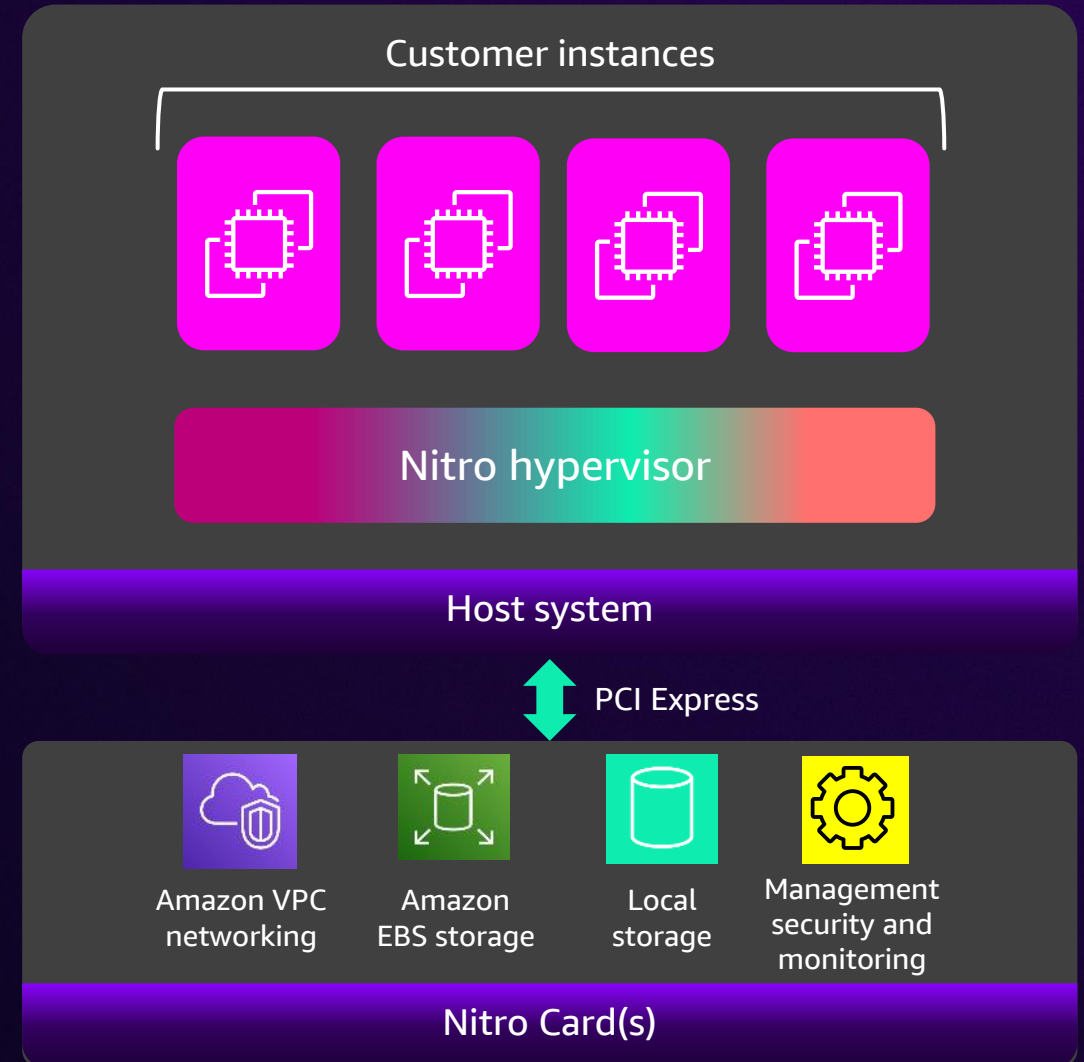
AWS Nitro Enclaves



Before Nitro System

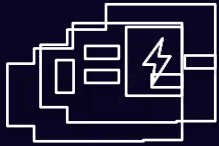


With Nitro System



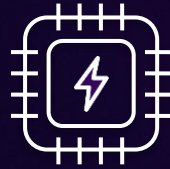
Nitro System core components

Nitro Cards



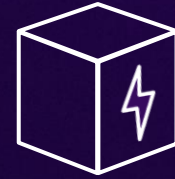
- Local NVMe storage
- Elastic block storage
- Networking, monitoring, and security

Nitro Security Chip



- Integrated into motherboard
- Protects hardware resources

Nitro Hypervisor



- Lightweight hypervisor
- Memory and CPU allocation
- Bare-metal-like performance

The Nitro System is the foundation for AWS

- All Amazon EC2 instance types released since 2018 are powered by the Nitro System
- Secure boot process based on a hardware root of trust ensures that every component is signed and every operation is pre-vetted for safety
- Every critical element of Nitro System can be live-updated
- Transparent encryption of storage, networking, and memory



Nitro-based Amazon EC2 server

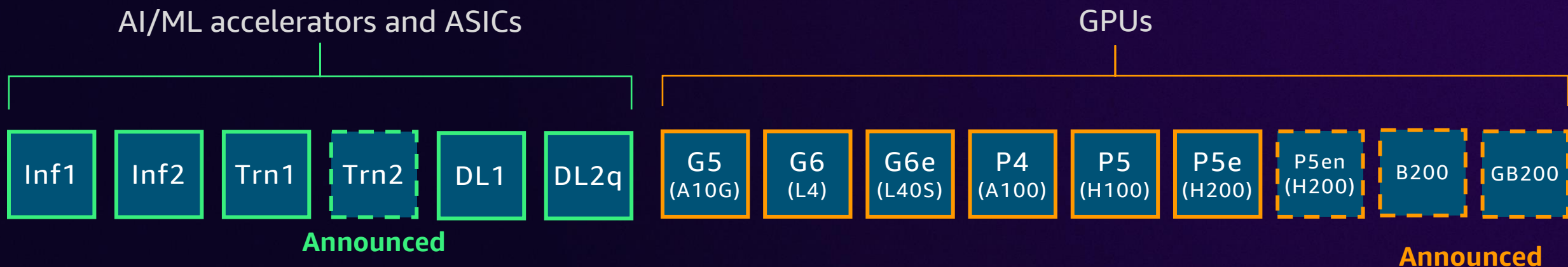
“Always-on” Confidential Computing

- There is **no operator access** mechanism in the Nitro System design
- No SSH or general purpose access of any kind
- All Nitro operations are done via secure, authenticated, authorized, logged (and audited) administrative APIs
- No APIs provide access to customer data



Nitro-based Amazon EC2 server

EC2 Accelerated Compute Instances for AI/ML



Trainium
Inferentia



H100, H200, B200,
GB200, A100, L40S,
A10G



Cloud AI100 Standard



Radeon GPU
Xilinx accelerator
Xilinx FPGA



Gaudi accelerator

Assurances and design information



[Whitepaper](#)

Security design of the
AWS Nitro System



[3rd party validation – Nitro Security](#)

NCC Group report - “no gaps in the
Nitro System that would
compromise these security claims”



[AWS Service Terms](#)

Updated service terms to
reflect “no operator access”

Nitro System key capabilities

NitroTPM



- Trusted Platform Module device attached to your EC2 instance
- Gather and attest an EC2 instance's state, store and generate cryptographic data, and prove platform identity

Nitro Enclaves



- Isolated compute environment
- Provides attestation with first class integration with AWS KMS

Attestation with Nitro Trusted Platform Module



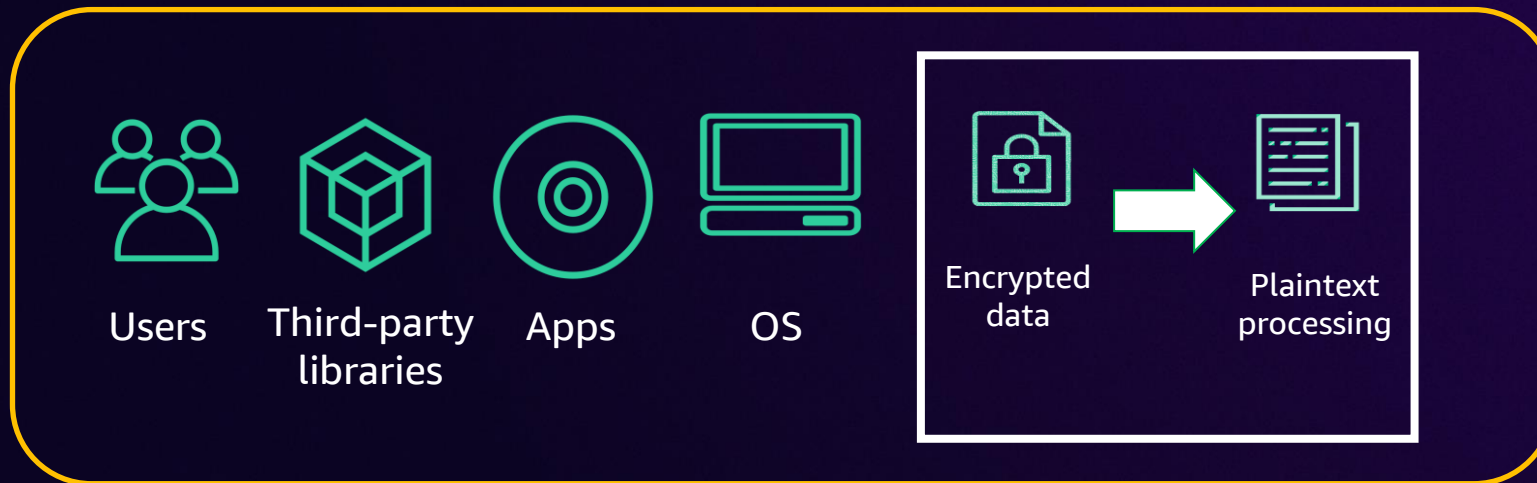
EC2 Instance

Proof of instance's state and identity



Attestation document

What is AWS Nitro Enclaves?



Amazon EC2 instance

What is AWS Nitro Enclaves?



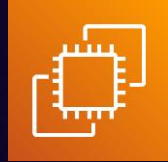
Nitro Enclaves provides additional isolation for data in use

AWS Nitro Enclaves features and benefits



Additional isolation and security

- No storage, access, or networking
- Secure local channel with parent EC2



Flexible

- Processor-agnostic technology
- Combinations of CPU and memory

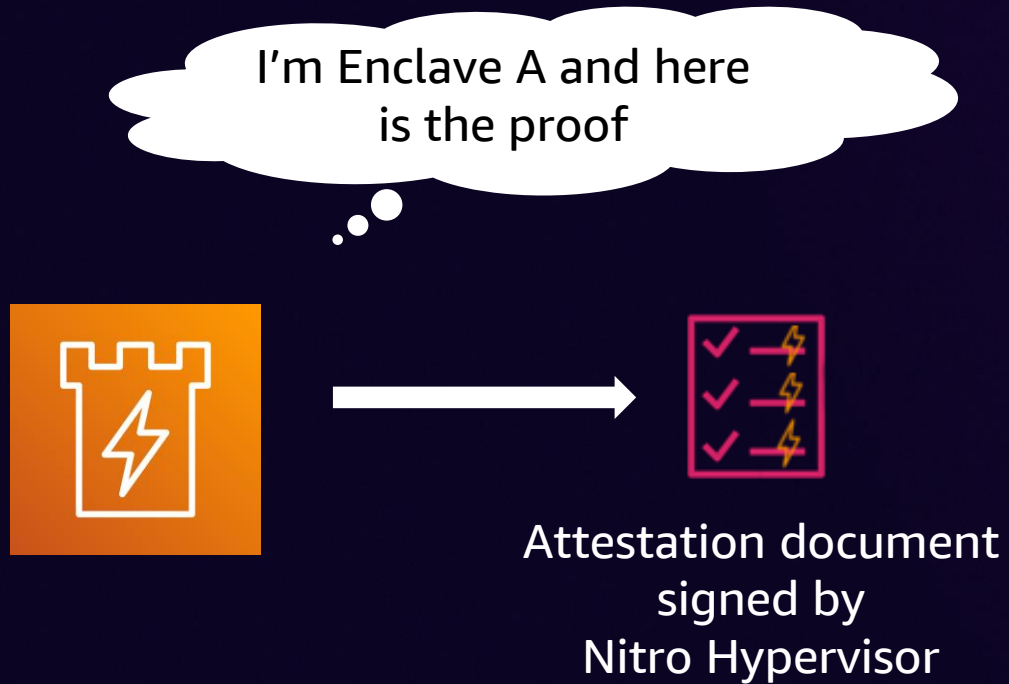


Cryptographic attestation

- Proves its identity and authorizes code
- Creates its own key pair and integrated with AWS KMS

No additional cost

Nitro Enclaves provides proof of identity



Attestation document – hashes

Enclave image

Application in the enclave

Parent instance ID

AWS Identity and Access Management (IAM)

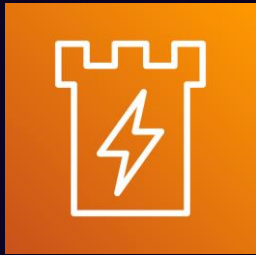
Role of parent instance

User defined info, e.g., Nonce

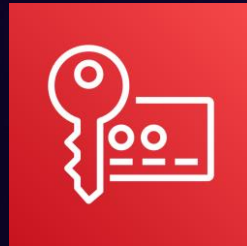
1st class integration with AWS KMS



Signed attestation
document



Nitro Enclaves

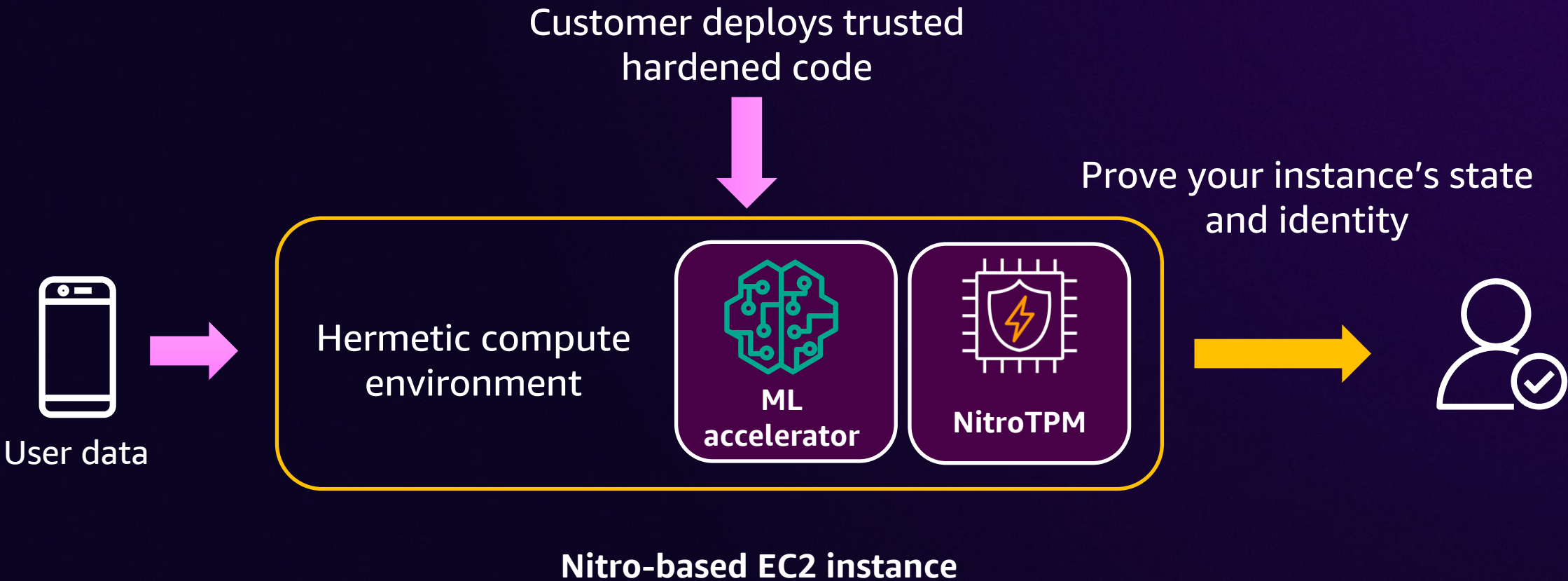


AWS KMS

- The **enclave** can prove its identity to **AWS KMS** with attestation document
- AWS KMS validates document against the AWS KMS key policy
- Enclave can perform cryptographic operations with AWS KMS keys

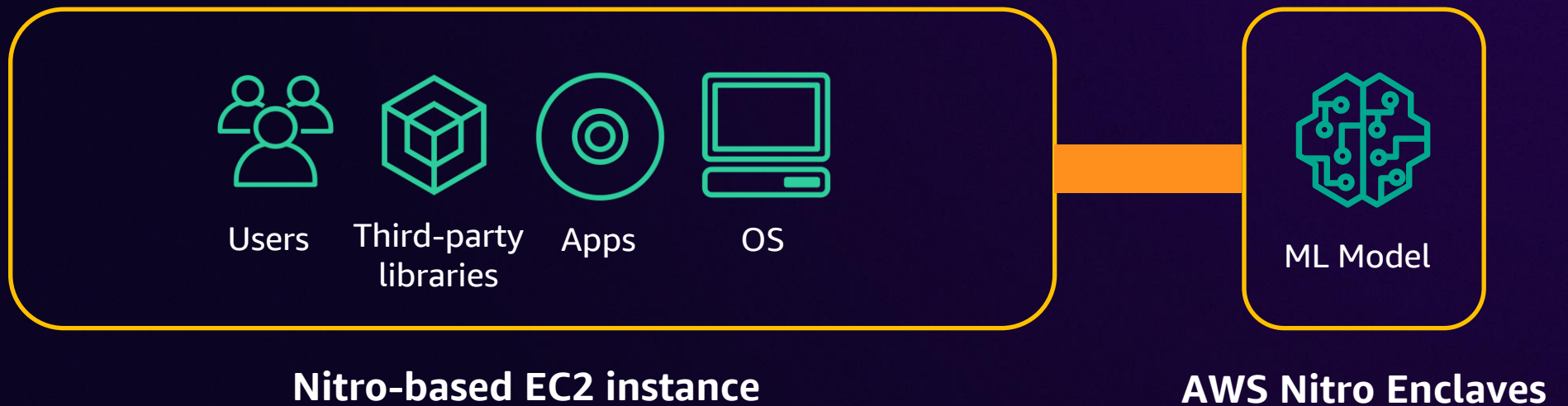
Confidential inference

ASSURANCE THAT USER DATA IS ONLY PROCESSED IN A TRUSTED ENVIRONMENT



End-to-end protection during inferencing

INTEGRATION SOLUTION BETWEEN NITRO ENCLAVES AND AWS KMS



Extending the Nitro System's end-to-end protection to ML accelerators

Coming Soon

AWS Nitro System

Encrypt your sensitive AI data using keys that you own and control

Store that data in a location of your choice

Securely transfer the encrypted data to an isolated compute environment for inferencing

Planned for the upcoming NVIDIA GB200 NVL72 and Trainium2



Tooling layer: Services for builders



Generative AI stack: Tooling layer

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

 **Amazon Bedrock**

Guardrails | Agents | Studio | Customization capabilities | Custom model import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



AWS Trainium



AWS Inferentia



Amazon SageMaker



Amazon EC2
UltraClusters



Elastic Fabric
Adapter



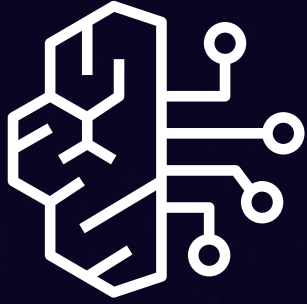
Amazon EC2
Capacity Blocks for ML



AWS
Nitro System



AWS
Neuron



Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)



Accelerate development of generative AI applications using FMs through an API, without managing infrastructure



Choose FMs from Amazon, AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, and Stability AI to find the right FM for your use case



Built with data security, privacy, and safety in mind

Amazon Bedrock

BROAD CHOICE OF MODELS

AI21labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI	stability.ai
Contextual answers, summarization, paraphrasing	Text summarization, generation, Q&A, search, image generation	Summarization, complex reasoning, writing, coding	Text generation, search, classification	Q&A and reading comprehension	Text summarization, text classification, text completion, code generation, Q&A	High-quality images and art
Jamba-Instruct Jurassic-2 Ultra Jurassic-2 Mid	Amazon Nova Micro, Lite, Pro Amazon Nova Canvas & Reel Amazon Titan Text Amazon Titan Embeddings Amazon Titan Multimodal Embeddings Amazon Titan Image Generator	Claude 3.5 Sonnet Claude 3 Opus Claude 3 Sonnet Claude 3 Haiku Claude 2.1 Claude 2 Claude Instant	Command Command Light Embed English Embed Multilingual Command R+ Command R	Llama 3.1 Llama 3 8B Llama 3 70B Llama 2 13B Llama 2 70B	Mistral Large 2 (24.07) Mistral Large (24.02) Mistral Small Mixtral 8x7B Mistral 7B	Stable Diffusion XL1.0 Stable Diffusion XL 0.8

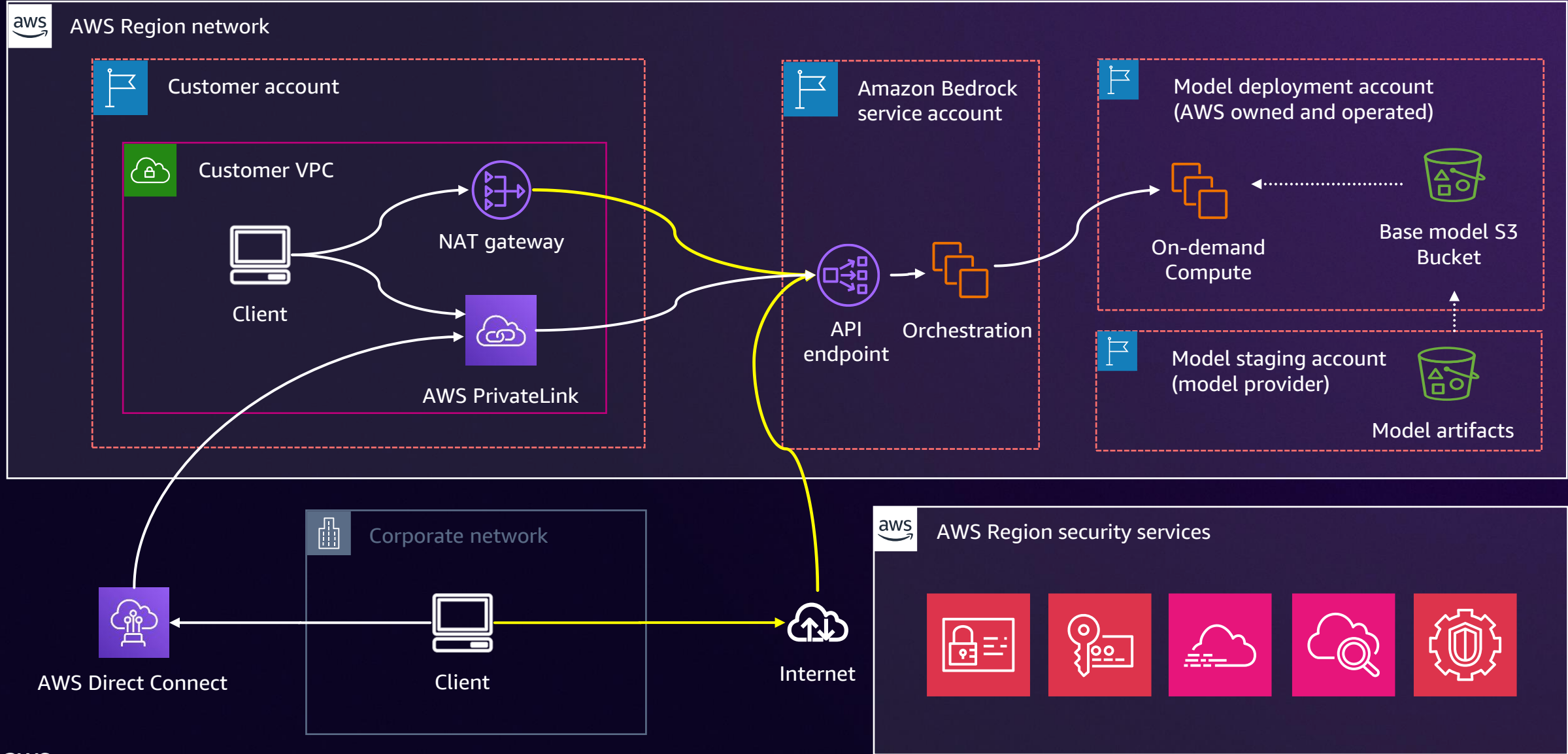
Amazon Bedrock

keeps data secure
and private

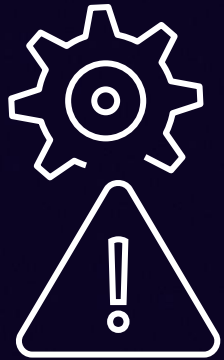


- None of the customer's data is used to train the underlying foundational models
- All data is encrypted at rest using AWS Key Management Service (KMS) and encrypted in transit with TLS 1.2 (minimum)
- Support for **AWS PrivateLink** so customers can establish private connectivity between virtual private clouds (VPCs) and the Amazon Bedrock service using VPC endpoints
- Fine-tuned models are encrypted and stored using customer AWS KMS key; only you have access to your customized models
- Data used to customize models remains within your VPC
- Integration with IAM for fine-grained access controls, and Amazon GuardDuty for threat detection
- Support for data privacy standards, including GDPR, HIPAA, and PCI

Amazon Bedrock data flows

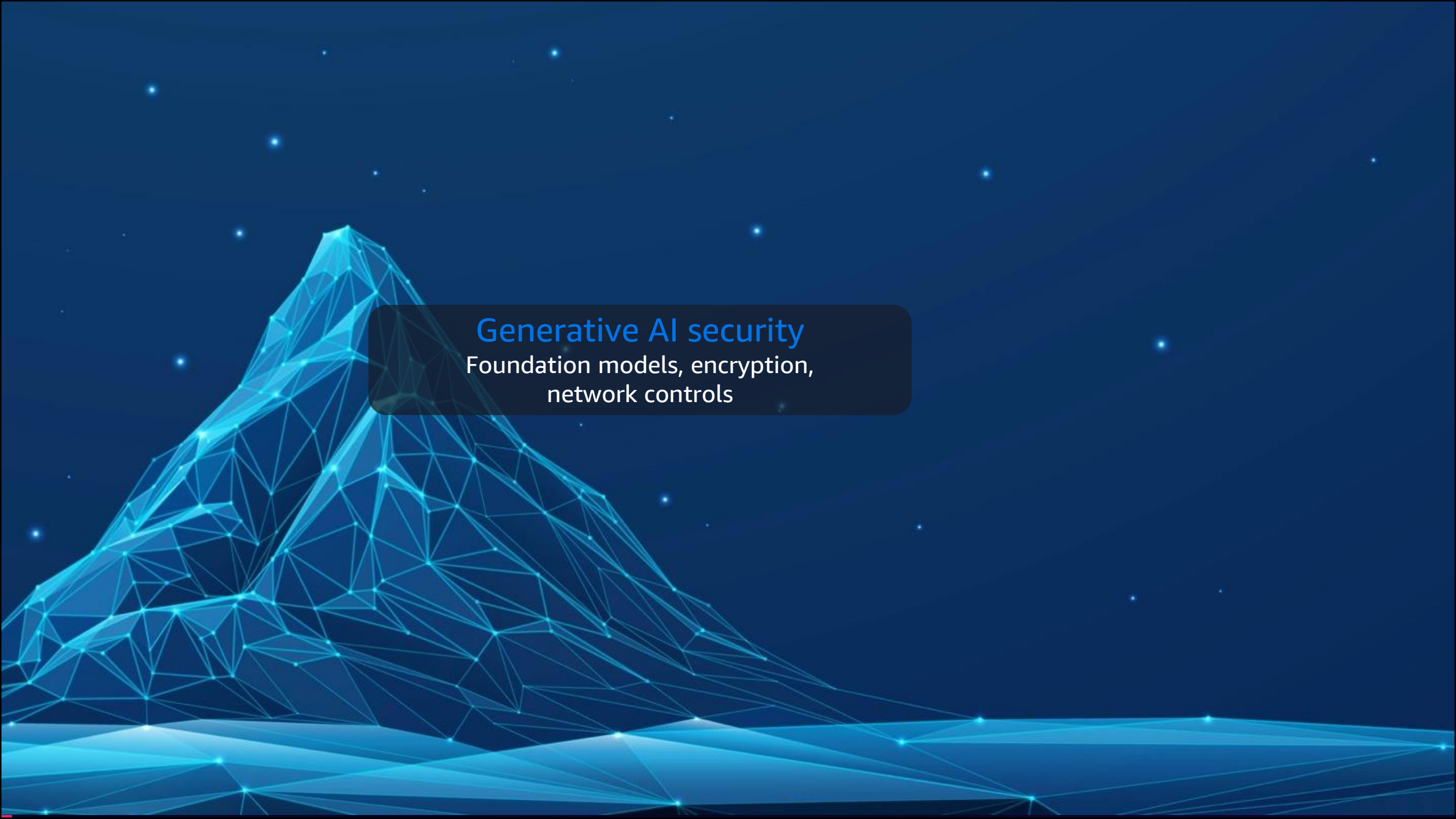


Governance and auditability support



Comprehensive monitoring and logging capabilities

- Track **usage metrics** and build customized dashboards using Amazon CloudWatch
- **Monitor API activity** and **troubleshoot issues** as you integrate other systems into your applications using AWS CloudTrail
- Compliance standards: C5, CISPE, DoD CC SRG IL2, ENS High, FINMA, **FedRAMP** (Moderate/High), **GDPR**, **HIPAA BAA**, ISMAP, **ISO** and CSA STAR, MTCS, OSPAR, **PCI**, Pinakes, PiTuKri, **SOC 1, 2, and 3**



Generative AI security

Foundation models, encryption,
network controls



Generative AI security

Foundation models, encryption,
network controls

Storage

Structured and unstructured data

Operational databases

SQL, NoSQL, document, graph, vector

Analytics and data lakes

Search, streaming, batch, interactive

Data integration

Capture, transformation, streaming

Data governance

Catalog, quality, privacy, access controls



Make generative AI
work with **your data**



RETRIEVAL AUGMENTED GENERATION (RAG)

Embed specialized knowledge through prompt augmentation

Use existing enterprise knowledge corpus

No change to the foundation model



AI AGENTS

Enable multi-step workflows using your enterprise systems

Take actions & query real-time data via APIs at inference time

No change to the foundation model



MODEL CUSTOMIZATION

Generalized and specialized knowledge for your domain

Fine-tuning and continued pre-training with enterprise data

Create a private copy of the foundation model

Limit access to your data, models, and outputs

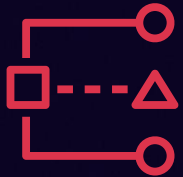
USE IDENTITY AND ZERO TRUST TO CONTROL ACCESS TO DATA IN GENERATIVE AI WORKLOADS



AWS IAM
Identity Center



Amazon Verified
Permissions



IAM
Access Analyzer

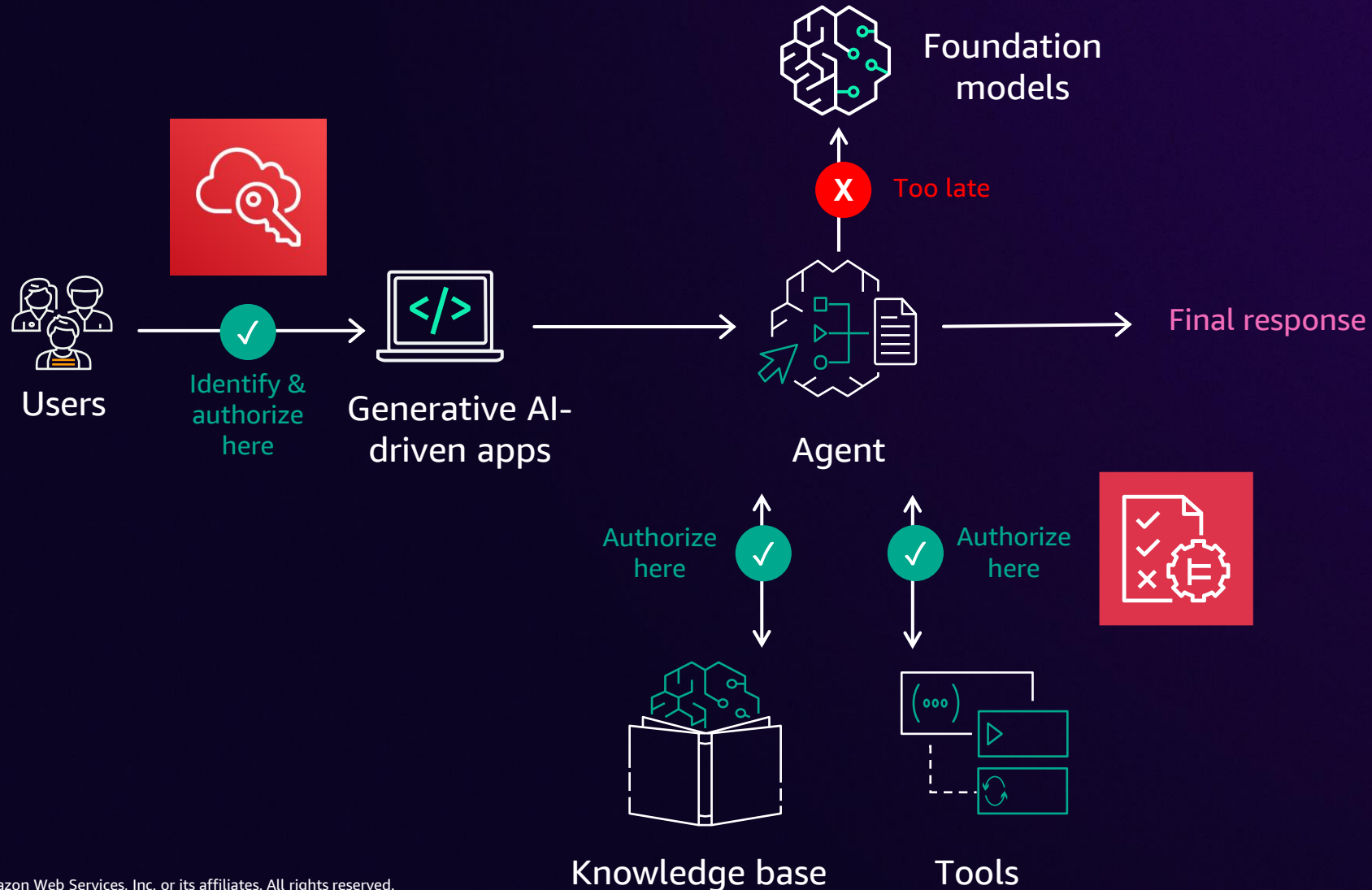


AWS Verified
Access

- Apply a policy of least privilege to training data, models, and applications using [AWS IAM Identity Center](#) and [IAM Access Analyzer](#).
- Explore further Zero Trust capabilities to add fine-grained access controls with [AWS Verified Access](#) and [Amazon Verified Permissions](#).
- Use [AWS Verified Access](#) to further eliminate the costs, complexity and performance issues related to VPNs.

Integrated security and identity

INTEGRATING TRADITIONAL SECURITY MECHANISMS IN YOUR GENERATIVE AI DEPLOYMENT



Amazon Bedrock Guardrails

Implement safeguards customized to your application requirements and aligned to your responsible AI policies

Blocks as much as 85% more harmful content than protection natively provided by some FMs on Amazon Bedrock today and filters over 75% of hallucinated responses for RAG and summarization workloads



Evaluate prompts and model responses for agents, knowledge bases, FMs in Amazon Bedrock, and custom or third-party FMs



Configure thresholds to filter harmful content, jailbreaks, and prompt injection attacks



Define and disallow denied topics with short natural language descriptions

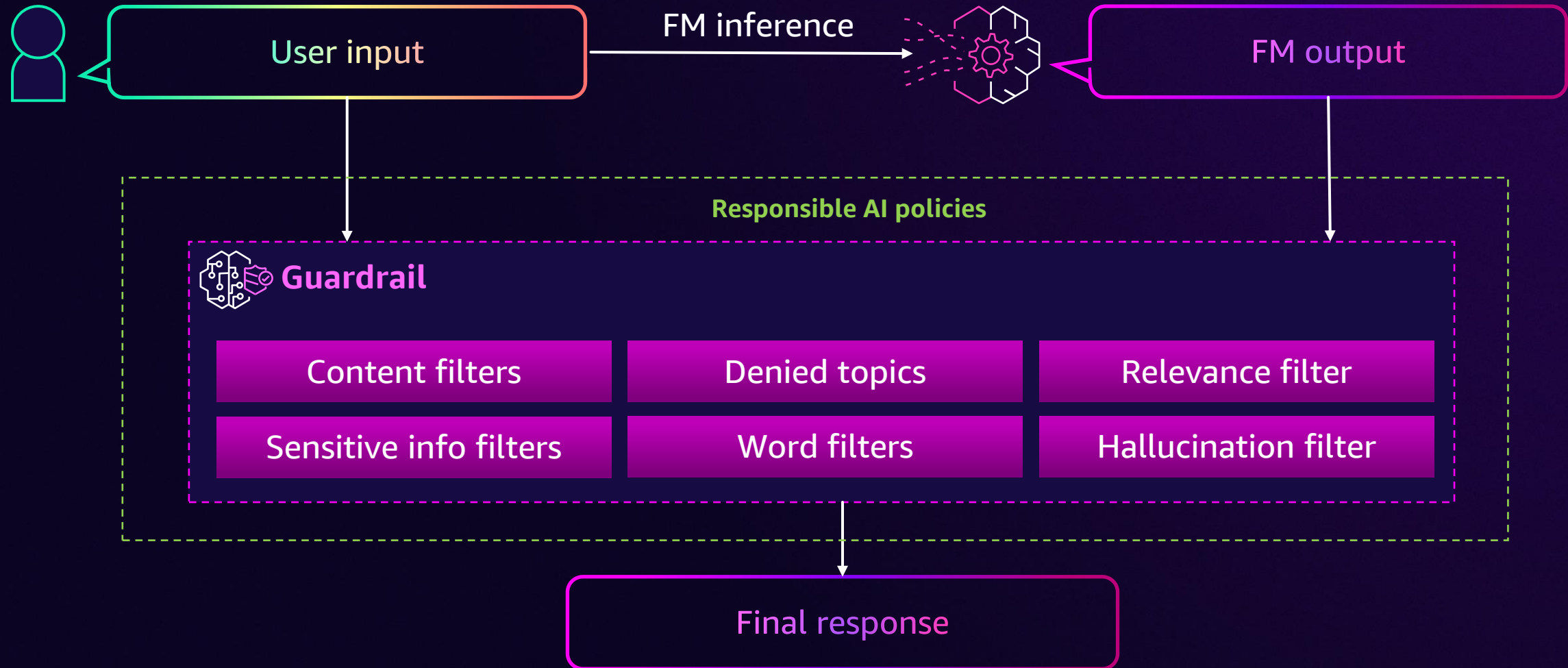


Remove personally identifiable information (PII) and sensitive information in gen AI apps



Filter hallucinations by detecting groundedness and relevance of model responses based on context

How it works: Amazon Bedrock Guardrails



Application layer: Enterprise-ready solutions



Generative AI stack: Enterprise applications

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs



Amazon Q
Business



Amazon Q
Developer



Amazon Q in
QuickSight



Amazon Q in
Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs



Amazon Bedrock

Guardrails | Agents | Studio | Customization capabilities | Custom model import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



AWS Trainium



AWS Inferentia



Amazon SageMaker



Amazon EC2
UltraClusters



Elastic Fabric
Adapter



Amazon EC2
Capacity Blocks for ML



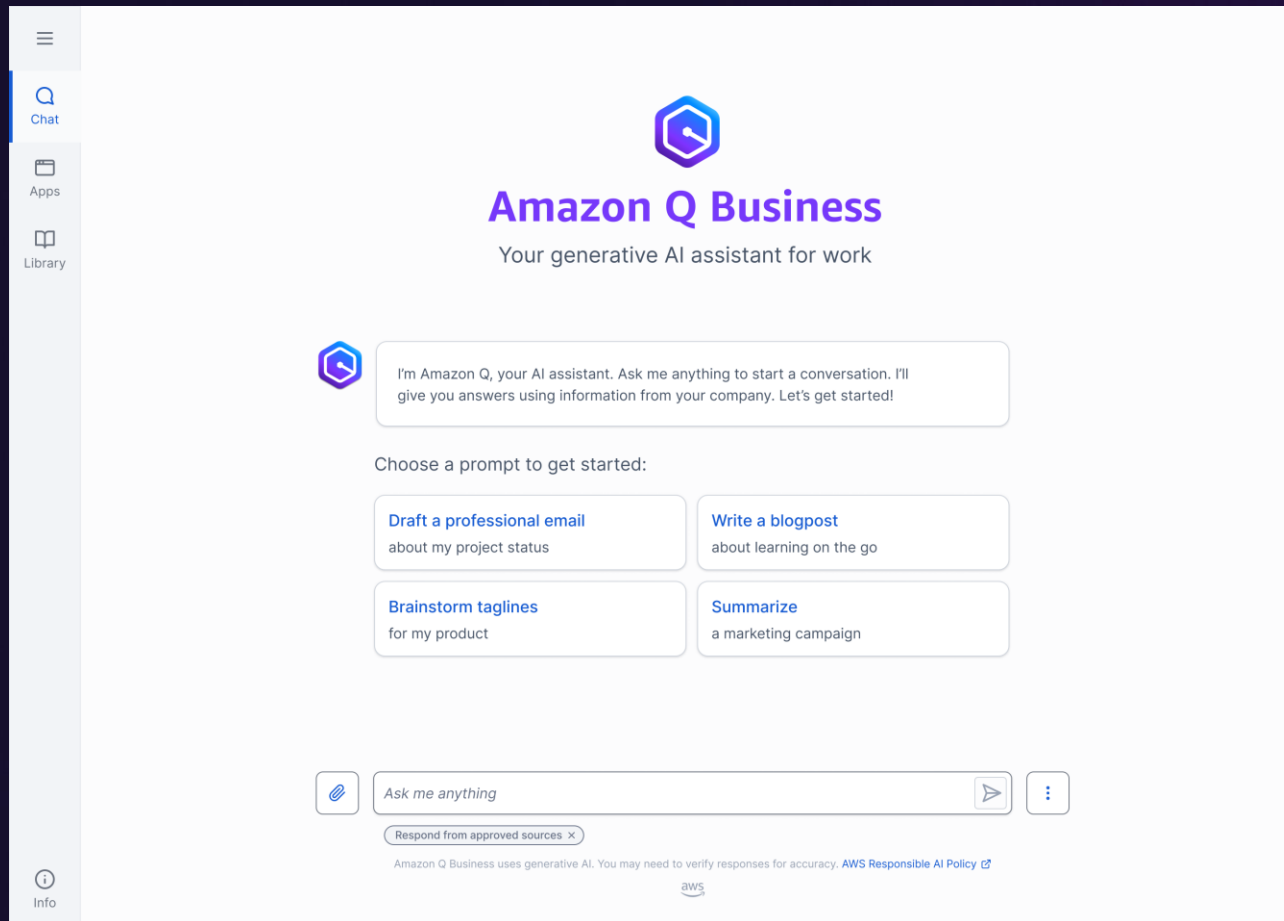
AWS
Nitro System



AWS
Neuron

Amazon Q Business overview

BOOST YOUR WORKFORCE PRODUCTIVITY WITH GENERATIVE AI



Delivers quick, accurate, and relevant answers to your business questions, securely and privately

Execute actions using out-of-the-box or custom plugins

Respects existing access control based on user permissions

Connects to over 40 popular enterprise applications and document repositories

Enables administrators to easily apply guardrails to customize and control responses

Streamlines daily tasks with user-created lightweight applications

Safety and security



Use pre-built guardrails for toxicity



Restrict responses to enterprise content only



Specify blocked words or phrases that never appear in responses



Define special topics and configure guardrails for such topics



Use AWS PrivateLink to access Amazon Q Business securely in your VPC environment using a VPC endpoint



Integrates with IAM Identity Center for easier management, auditability, and control



Integrates with AWS CloudTrail to record actions taken by a user, role, or an AWS service

FIPS

Support for Federal Information Processing Standard (FIPS) endpoints

Data privacy and governance

You are always in control of your data



- Amazon Q Business uses pre-trained machine learning models
- Customer data (i.e., data ingested into Amazon Q Business, conversation data, as well as feedback data) is not used for training, fine-tuning, or improving machine learning models by AWS
- Amazon Q Business is a regional service: Customer data is stored in the region where the Amazon Q Business App is created
- Customer data is always encrypted in transit with a minimum of TLS1.2
- Amazon Q Business encrypts all the data at rest using [AWS Key Management Service \(KMS\)](#)

Find accurate and referenceable answers



Trusted answers generated from enterprise data



In-context conversations



Source references for fact-checking



Conversation history



Tune the relevance of results with index boosting

data durability, which means the data is redundantly stored across multiple Availability Zones that are physically separated within an AWS Region . [1] [2] The source also states that S3 is resilient to AZ failures and stores data across multiple AZs, resulting in high data durability . [3] Additionally, the Resilience Hub assessment shows that S3 does well in terms of resiliency, with the S3 bucket being resilient to AZ failures . [3]

Sources ▾

👤 Tell me more about availability zones

🔍 Availability Zones are distinct locations within an AWS Region that are engineered to be isolated from failures in other Availability Zones. [1] [2] Each Availability Zone provides inexpensive, low-latency network connectivity to other Availability Zones in the same AWS Region. [1] [2] By launching instances in separate Availability Zones, you can protect your applications from the failure of a single location. [1] [2] AWS maintains multiple Availability Zones in each Region, and these Availability Zones are physically isolated from each other but united by private, low-latency, high-throughput, and highly redundant network connections. [3] [4] This enables AWS to provide very high levels of availability and redundancy, while also minimizing latency. [3] [4]

Sources ▾

👍 🗨️ 📄

🔗 Enter a prompt

Responding from approved sources

Amazon Q Business uses generative AI. You may need to verify responses for accuracy. [AWS Responsible AI Policy](#)

aws

Custom plugins

The screenshot displays the Amazon Q Business interface. At the top left is a blue menu icon. Below it, a user profile icon is followed by the text "I'd like to request time off". A purple hexagonal icon with a white 'Q' is positioned to the left of a white form box. Inside the form box, the text "You are connected to HR Time Off." is displayed, followed by "Request your time off." Below this, there is a "Time off type" dropdown menu with "Paid Time Off (PTO)" selected. Underneath are "From" and "To" date input fields, with "07/01/2024" and "07/05/2024" respectively. At the bottom right of the form are "Cancel" and "Create" buttons. Below the form is a text input field with the placeholder "Ask me anything" and a "Send" button. A small "HR Time Off x" tag is visible below the input field. At the bottom of the interface, there is a footer with the text "Amazon Q Business uses generative AI. You may need to verify responses for accuracy. AWS Responsible AI Policy" and the AWS logo.

Conversation settings

- Use a plugin
- HR Time Off
- Jira
- Respond from approved sources

Use custom plugins to access and update information in enterprise systems such as Vanguard, ADP, Dynamics 365, or Microsoft Exchange using natural language queries in Amazon Q Business to enable tasks such as scheduling meetings, capturing meeting notes, getting or updating sales information, and more!

Putting it all together



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Defense-in-depth security

LAYERED SECURITY CONTROLS FOR GENERATIVE AI

Policies, procedures and awareness

Network and edge protection

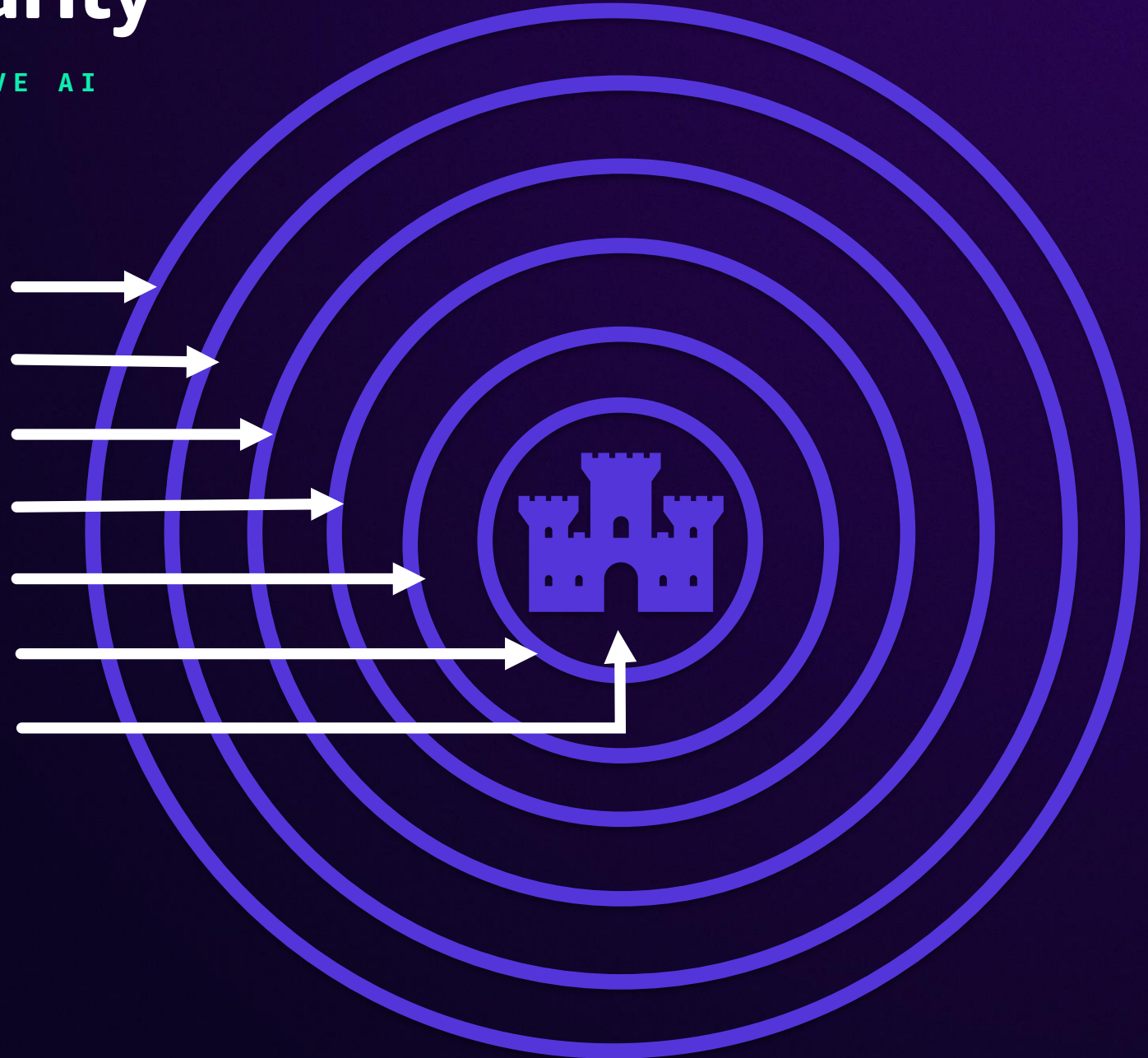
Identity and access management

Threat detection and incident response

Infrastructure protection

Application protection

Data protection



AWS generative AI and security integrated together

FOUNDATIONAL AWS SECURITY + ADDITIONAL SECURITY FEATURES OF GENERATIVE AI SERVICES

AWS generative AI services



Amazon
Bedrock



Amazon
SageMaker



Amazon Q
Business



Amazon Q
Developer



Amazon
CodeGuru Security

AWS security, identity, and compliance services



AWS
Security Hub



AWS
KMS



Amazon
GuardDuty



AWS
Shield
Advanced



AWS
WAF



AWS
Network
Firewall



AWS Audit
Manager



Amazon
Macie



Amazon
Inspector



Amazon
Detective



AWS IAM
Identity
Center



AWS IAM
Access
Analyzer



Amazon
Verified
Permissions



AWS Artifact

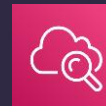


AWS Signer

AWS cloud ops, networking, and storage



AWS
CloudTrail



Amazon
CloudWatch



AWS Systems
Manager



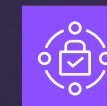
AWS
Config



AWS
Trusted
Advisor



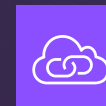
AWS Well-
Architected
Tool



AWS
Verified
Access



Amazon
VPC



AWS
PrivateLink



Amazon
S3 Object
Lock



AWS
Backup

Resources



Landing page:
Securing generative AI
on AWS



Blog post:
A secure approach to
generative AI with AWS



Blog post:
Data authorization in
generative AI apps

Thank you!

Jason Garman

garmaja@amazon.com

JD Bean

jdbean@amazon.com



Please complete the session
survey in the mobile app