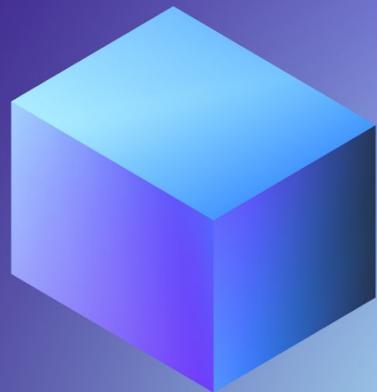




Generative AI with AWS

Rafa Gomez

Senior Solution Architect



2023

The Year of POCs



What is generative AI?

Is this secure?

Do I need to become a prompt engineer?

How do I choose a model?

Where do I get started?



What does this mean for my business?

What is a Foundation Model?



Which models should we try out?

What is FM?

What is a Large Language Model?

2024

The Year of Production

(FOR SOME)



How do I prioritize my projects?

How can I lower my costs?

How do I make this real?

What customization method should I use?



How I can I scale this?

Which models should I use?

Should I train my own model?

How do I manage risks?



How can we move faster?

Generative AI has potential to create significant business value



NEW EXPERIENCES

Create new innovative and engaging ways of interacting with your customers and employees



PRODUCTIVITY

Radically improve productivity across all lines of business



INSIGHTS

Extract insights and clear answers from all your corporate information, enabling faster and better decisions



CREATIVITY

Create new content and ideas, including conversations, stories, images, videos, and music



Generative AI Application



Generative AI
Application

Data Foundation

STORAGE

GOVERNANCE
& COMPLIANCE

DATABASES,
ANALYTICS,
& DATA LAKES

DATA INTEGRATION

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPU
s



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity
Blocks



Nitro



Neuron



Amazon SageMaker

Build, train, and deploy ML models at scale, including FMs

Access the latest and publicly available FMs

Build FMs from scratch

Customize FMs

Run inference

Implement FMOps and governance

Innovating at the silicon level

AWS Trainium



AWS Inferentia

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs



Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPU
s



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity
Blocks



Nitro



Neuron

Amazon Bedrock

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

Choice of leading FMs through a single API

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and safety



Amazon Bedrock Helps keep your data secure and private



None of the customer's data is used to train the underlying model

All data is encrypted in transit and at rest; data used for customization is securely transferred through customer's VPC

Data remains in the Region where the API is processed

Support for GDPR, SOC, ISO, CSA compliance, and HIPAA eligibility

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 | Customization Capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPU
S



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity
Blocks



Nitro



Neuron

Amazon Q



AMAZON Q BUSINESS

AMAZON Q DEVELOPER

EMBEDDED

**Amazon Q
In Connect**

**Amazon Q
In QuickSight**

**Amazon Q
In AWS Supply Chain**





Thank you!

Rafa Gomez

Senior Solution Architect