CAYLENT

Measuring the Performance of Serverless Applications with Generative Models in Amazon Bedrock

Hazel Sáenz - AWS Serverless Hero - Cloud Software Architect

About me!

Hazel Sáenz

I'm a Software Architect in Caylent with over a five years of experience designing scalable cloud-native applications. As an AWS Serverless Hero, I've led complex cloud migration and system reengineering projects.

I'm passionate about serverless technologies and DevOps, and I frequently share my experience at AWS events across Latin America.

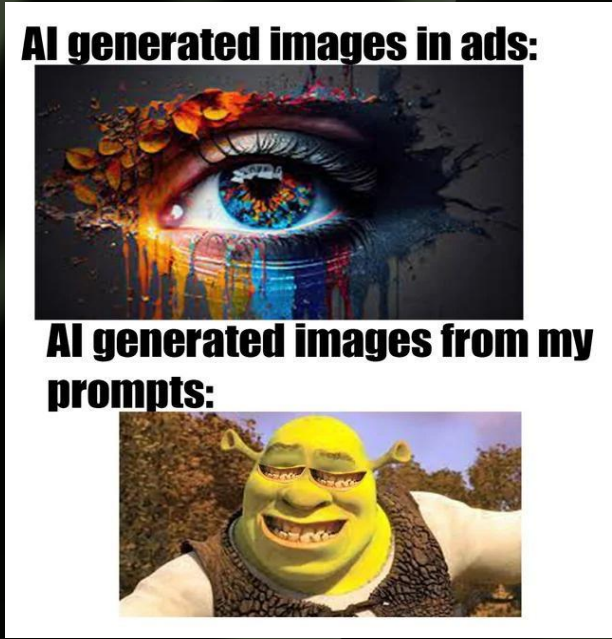
I'm also the creator of Kiu, a virtual agent powered by serverless and generative AI, designed to transform how we learn about the cloud.

aws serverless HERO

https://hazelsaenz.tech

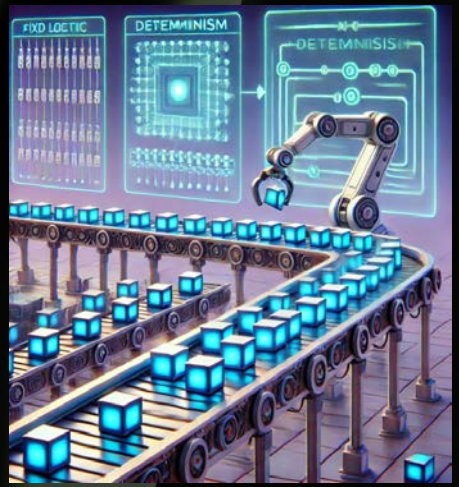
CAYLENT

What are you going to learn today?



"Building an ai App that never fails...
sounds easy, doesn't it?"

Why is so hard?



Deterministic

Probabilistic

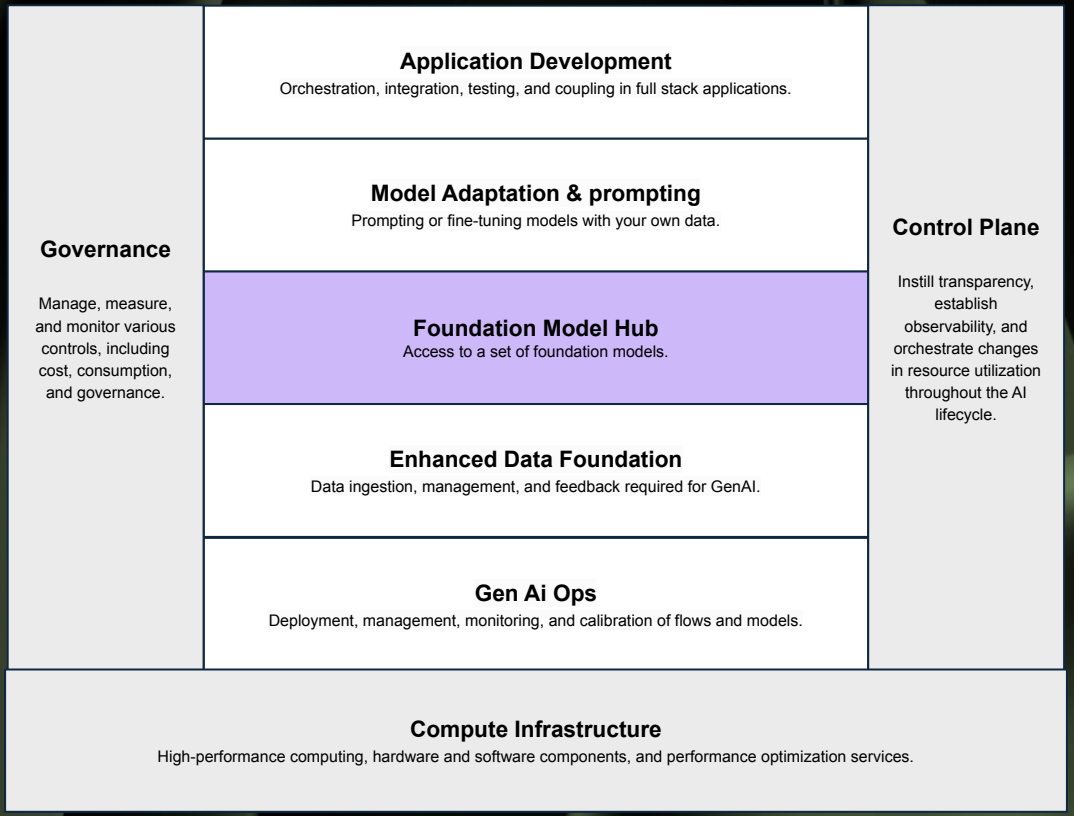
CAYLENT

What is the difference between them?

Feature	Deterministic	Probabilistic
Fixed result	✅ Yes	❌ No
Depends on fixed rules	✅ Yes	❌ No (depends on data and models)
Repeatability	✅ Always the same	❌ May vary
Example in AI	Classical algorithms (e.g., binary search)	ML models like Amazon Nova, Claude, Stable Diffusion

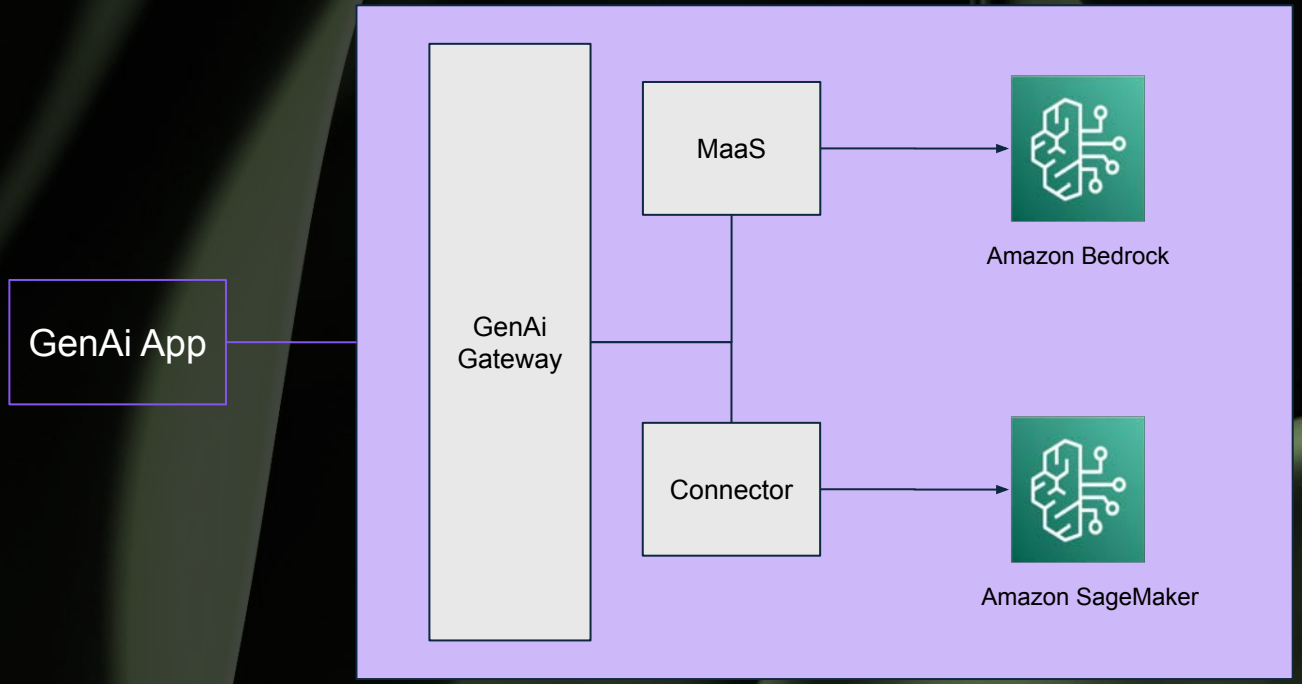
GenAI Objectives

Building Blocks



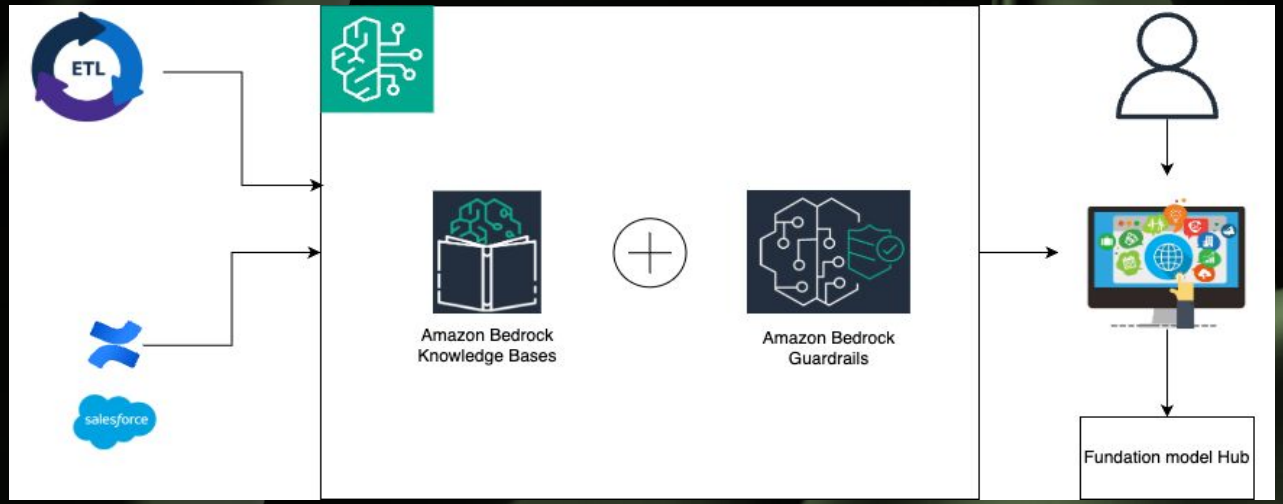
CAYLENT

Foundation Model Hub

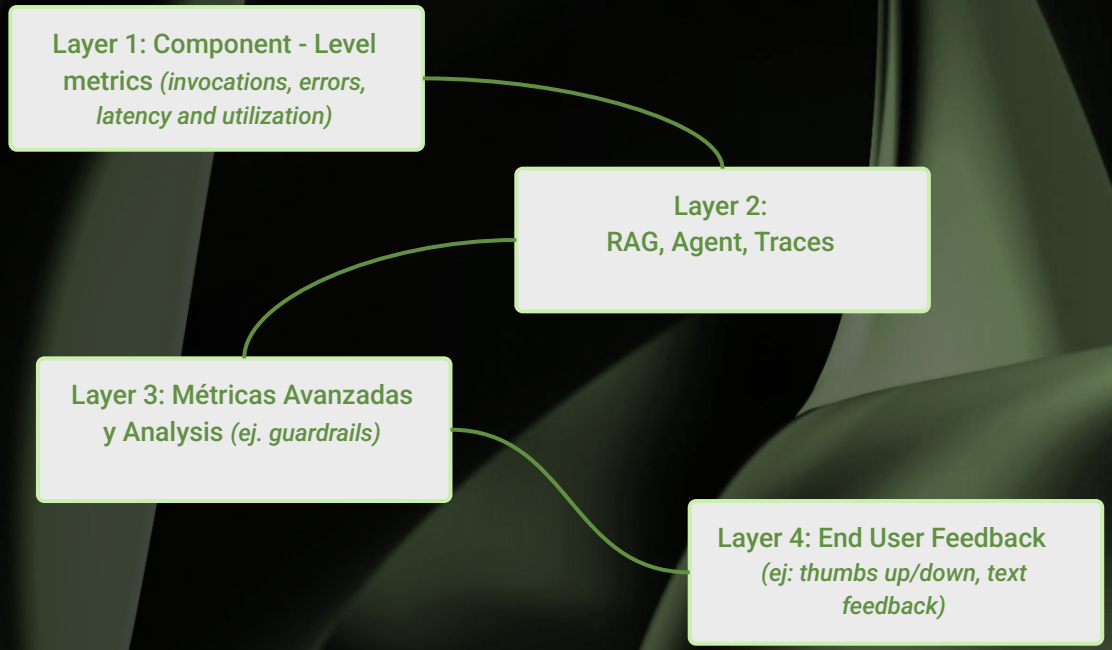


CAYLENT

Enhanced Data Foundation

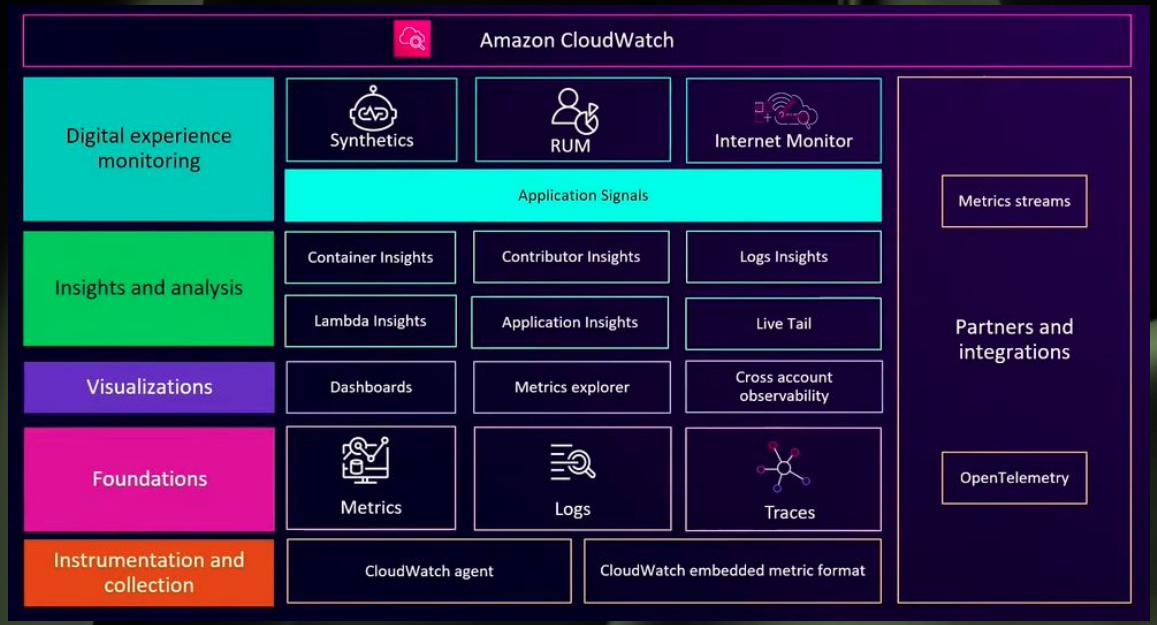


Layers to implement Observability



CAYLENT

Amazon CloudWatch



CAYLENT

Demo

CAYLENT

Información general Información

Modelos fundacionales

Amazon Bedrock supports over 100 foundation models from industry-leading providers and emerging leaders. Select a serverless model or Bedrock Marketplace model that is best suited for achieving your unique goa

View Model catalog | Discover marketplace models

Model spotlight

Anthropic's Claude

Choose the exact combination of intelligence, speed, and cost to suit your needs. All of the latest Claude models, like upgraded Claude 3.5 Sonnet, are available in Amazon Bedrock.

Solicitar acceso al modelo

Chat / Text

Generate text for a vast range of language processing tasks with various pre-trained models. You can use a single prompt or iterate on the result by submitting subsequent prompts that take into account the context of previous prompts and generated responses in a chat format.

Open playground

Image / Video

Genere fácilmente imágenes atractivas proporcionando solicitudes de texto a modelos previamente en
En el área de juego, introduzca un texto para comenzar.

Área de juego de imagen abierta

Orquestaciones

Create GenAI applications, augment responses with your proprietary data, and experiement with prompts.

Prompt Management

Create, test, and manage prompts and configure models and inference parameters used to run them.

View Prompt Management

Bases de conocimientos

Utilice orígenes de datos para aumentar los modelos y generar respuestas precisas y específicas para cada contexto. Puede agregar estos orígenes de datos para crear aplicaciones para muchos casos de uso, como la búsqueda semántica y la clasificación.

Agentes

Integre agentes para acelerar la entrega de aplicaciones de IA generativa utilizando las capacidades de razonamiento avanza los modelos básicos para procesar las solicitudes de los usuari
Después de crear un agente, puede probarlo en tiempo real.