Cloud Computing para Inteligencia Artificial

Sesión 14: Casos de Estudio, Arquitecturas y Tendencias Futuras

Grado en Ingeniería en Inteligencia Artificial

Agenda de la Sesión

1

Arquitecturas de Referencia (70 min)

- Caso 1: Sistema de Recomendación a Recomendación a Gran Escala
- Caso 2: Análisis de Imágenes Médicas
- Caso 3: Análisis de Sentimiento en en Redes Sociales
- Caso 4: Detección de Fraude en Tiempo Real

2

Desafíos y Tendencias Futuras (40 min)

- Desafíos Comunes en la Implementación
- El Futuro: IA Generativa, Edge AI, AI, Serverless, Hardware y más

3

Cierre y Discusión Abierta Abierta (10 min)

Repaso general del curso y Q&A

En esta sesión integraremos todo lo aprendido para diseñar soluciones completas de IA en la nube.

Caso 1: Sistema de Recomendación Personalizado

Objetivo

Ofrecer recomendaciones de productos o contenidos en tiempo real a millones de usuarios.

Componentes Clave de la Arquitectura

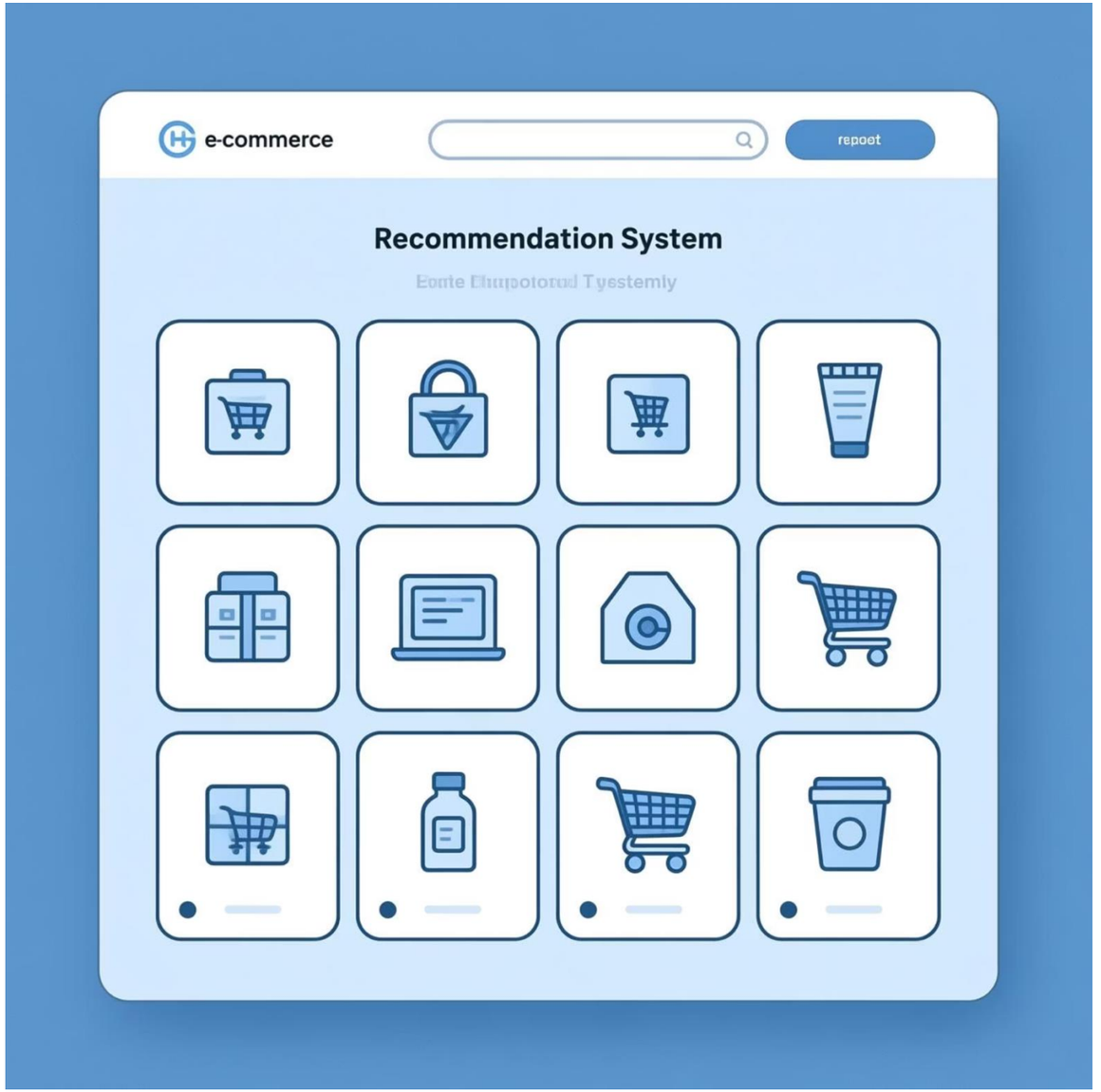
Ingesta de Datos: Captura de interacciones de usuario (clics, vistas, vistas, compras)

Almacenamiento: Persistencia de perfiles de usuario y catálogo de ítems de ítems

Entrenamiento del Modelo: Procesamiento de datos y entrenamiento (filtrado colaborativo, basado en contenido, híbrido)

Inferencia en Tiempo Real: API de baja latencia que devuelve recomendaciones

Ciclo de Feedback: Reincorporación de nuevas interacciones para para reentrenar el modelo



Arquitectura de Referencia en AWS (Recomendación)

Ingesta

Amazon Kinesis Data Streams/Firehose para capturar eventos en tiempo real

Almacenamiento

S3 Data Lake para datos crudos y procesados

DynamoDB para perfiles de usuario y catálogo (acceso de baja latencia)

Entrenamiento

Amazon SageMaker usando algoritmos integrados (Factorization Machines) o trayendo tu propio modelo (BYOM)

Inferencia

SageMaker Endpoints como API autoescalable

Solución Gestionada: **Amazon Personalize** abstrae gran parte de esta arquitectura

Alternativas en GCP y Azure (Recomendación)

Google Cloud Platform (GCP)

- **Ingesta:** Pub/Sub
- **Almacenamiento:** Cloud Storage, Storage, Bigtable/Firestore
- **ML:** Vertex AI Training/Prediction
- **Servicio Gestionado:** Recommendations AI

Microsoft Azure

- **Ingesta:** Event Hubs
- **Almacenamiento:** Blob Storage, Cosmos DB
- **ML:** Azure Machine Learning
- **Servicio Gestionado:** Azure AI Personalizer

El concepto clave es entender los bloques funcionales, no solo memorizar los nombres de los servicios. Si comprendes que necesitas un "message broker", un "object store" y una "plataforma de ML", podrás diseñar en cualquier nube.

Caso 2: Análisis de Imágenes Médicas (Detección de Anomalías)

Objetivo

Crear un sistema que analice imágenes médicas (Rayos X, resonancias) para ayudar para ayudar a los radiólogos a detectar anomalías.

Consideraciones Críticas

Seguridad, privacidad y cumplimiento normativo (HIPAA)

Ingesta Segura

Transferencia encriptada de archivos (formato DICOM)

Entrenamiento de Modelo

Uso de CNNs en hardware especializado (GPUs)

Almacenamiento Conforme

Control de acceso estricto, encriptación y auditoría

Inferencia y Revisión Humana

API para predicción y una interfaz para validación por expertos (Human-in-the-loop)

Preprocesamiento

Normalización de imágenes, aumento de datos

Arquitectura de Referencia en AWS (Imágenes Médicas)

Ingesta/Transferencia

AWS DataSync o Storage Gateway para mover datos desde el
el hospital

Almacenamiento

S3 con encriptación (SSE-S3/KMS), políticas de IAM/Bucket Policies
Policies estrictas y versionado

Entrenamiento

SageMaker con instancias de GPU (P3, G4), usando frameworks
frameworks como TensorFlow o PyTorch

Inferencia

SageMaker Endpoints para predicciones en tiempo real

Revisión Humana

Amazon A2I (Augmented AI) para integrar flujos de revisión

Auditoría

AWS CloudTrail y CloudWatch para registrar cada acceso y acción

Alternativas en GCP y Azure (Imágenes Médicas)

Google Cloud Platform (GCP)

- **Servicios Específicos:** Cloud Healthcare API para manejar datos DICOM de DICOM de forma nativa
- **Almacenamiento:** Cloud Storage con controles de acceso
- **ML:** Vertex AI Training/Prediction

Microsoft Azure

- **Servicios Específicos:** Azure Health Data Services (soporte para FHIR/DICOM)
- **Almacenamiento:** Blob Storage con controles de acceso
- **ML:** Azure Machine Learning

Tanto GCP como Azure han desarrollado APIs específicas para el sector salud, mientras AWS ofrece bloques de construcción genéricos y seguros.

Caso 3: Análisis de Sentimiento en Redes Sociales

Objetivo

Monitorizar menciones de una marca en redes sociales y clasificarlas como positivas, negativas o neutras en tiempo real.

Ingesta de Datos

Conexión a APIs de redes sociales para obtener un stream de publicaciones

Procesamiento/Análisis

Limpieza del texto y clasificación del sentimiento

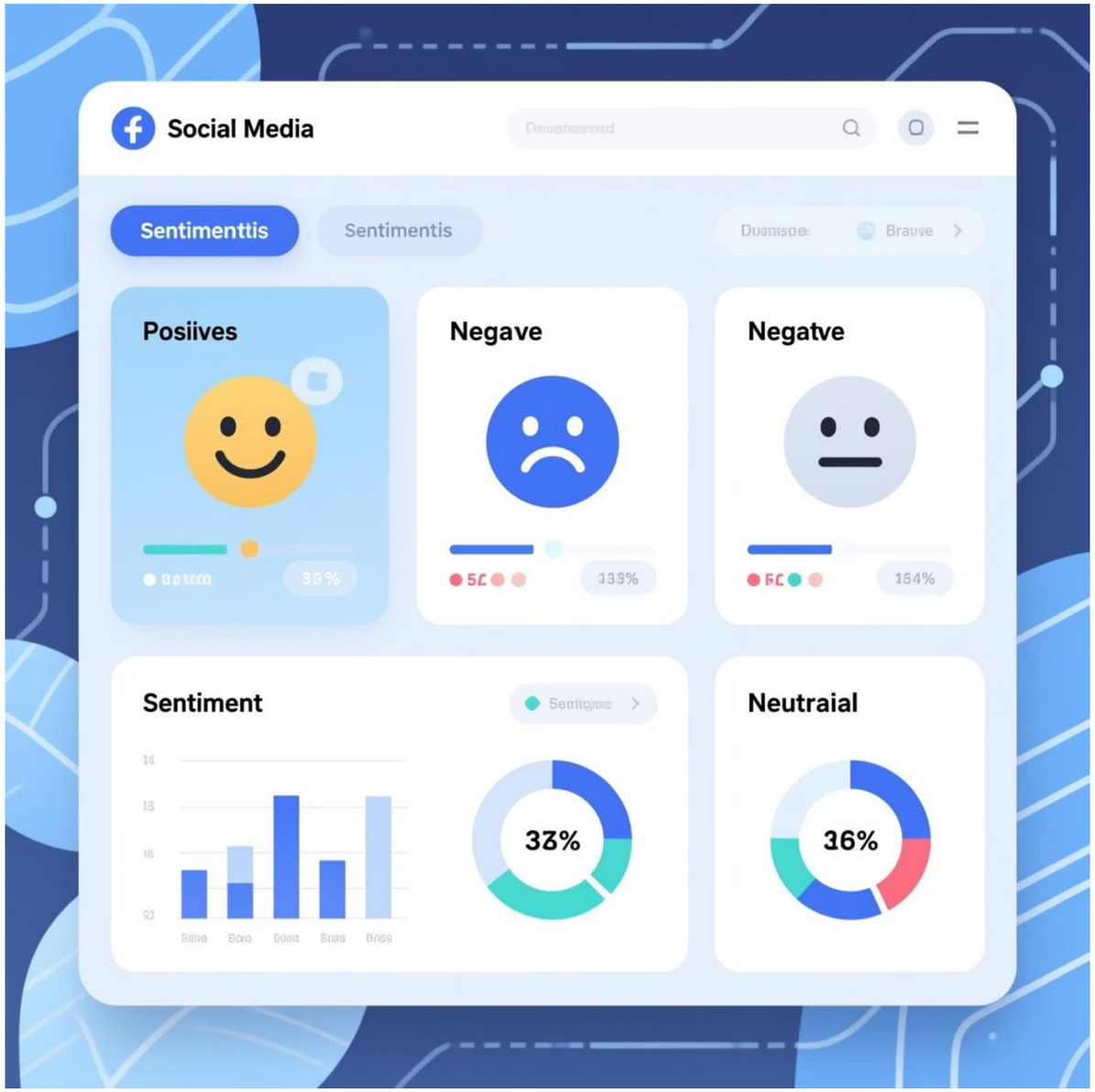
Almacenamiento

Guardar los datos de texto crudos

Visualización

Creación de dashboards para monitorizar los resultados

Decisión Clave: ¿Usar una API pre-entrenada o entrenar un modelo personalizado?



Arquitectura de Referencia en AWS
AWS (Análisis de Sentimiento)

Ingesta

Lambda functions que consultan las APIs y envían los datos a **Kinesis Firehose**

Almacenamiento

Kinesis Firehose guarda los datos directamente en **S3**

Procesamiento/Análisis

Opción 1: Lambda que llama a Amazon Comprehend (API de NLP)

Opción 2: SageMaker (con BlazingText o modelo de Hugging Face)

Visualización

Amazon OpenSearch Service (con Dashboards) o QuickSight

Alternativas en GCP y Azure (Análisis (Análisis de Sentimiento)

Google Cloud Platform (GCP)

- **Ingesta:** Cloud Functions + Pub/Sub
- **Análisis Gestionado:** Natural Language Language API
- **Análisis Personalizado:** Vertex AI (AutoML Text o Custom Training)
- **Visualización:** BigQuery + Looker Studio

Microsoft Azure

- **Ingesta:** Azure Functions + Event Hubs
- **Análisis Gestionado:** Azure AI Language Language (antes Text Analytics)
- **Análisis Personalizado:** Azure Machine Machine Learning
- **Visualización:** Azure Synapse Analytics + Power BI

Cada proveedor ofrece una API de lenguaje pre-entrenada y su plataforma de ML para construir modelos a medida. Los ecosistemas tienden a integrarse con sus propias herramientas de Business Intelligence.

Caso 4: Detección de Fraude en Transacciones Financieras

Objetivo

Analizar transacciones con tarjeta de crédito en tiempo real y bloquear las potencialmente fraudulentas.

Consideraciones Críticas

- Latencia ultra baja (decenas de milisegundos)
- Alta disponibilidad y reentrenamiento continuo del modelo

01

Ingesta de Transacciones

Stream de eventos de alta velocidad

02

Ingeniería de Características

Enriquecimiento de datos en tiempo real (ej. ¿cuántas compras ha hecho el usuario en la última hora?)

03

Motor de Inferencia

Modelo de ML optimizado para una latencia mínima

04

Sistema de Alertas

Notificación inmediata si se detecta fraude

05

Reentrenamiento Continuo

El modelo debe adaptarse a los nuevos patrones de fraude

Arquitectura de Referencia en AWS (Detección de Fraude)

Ingesta

Kinesis Data Streams para un stream de stream de baja latencia

Ingeniería de Características

AWS Lambda o Kinesis Data Analytics (con Flink) para procesar eventos sobre la marcha

Inferencia

SageMaker Endpoint con modelo optimizado (XGBoost compilado con con Neo)

O Amazon Fraud Detector como servicio gestionado

Alertas

El sistema publica en Amazon SNS para enviar alertas

Reentrenamiento

SageMaker Pipelines para orquestar el reentrenamiento automático

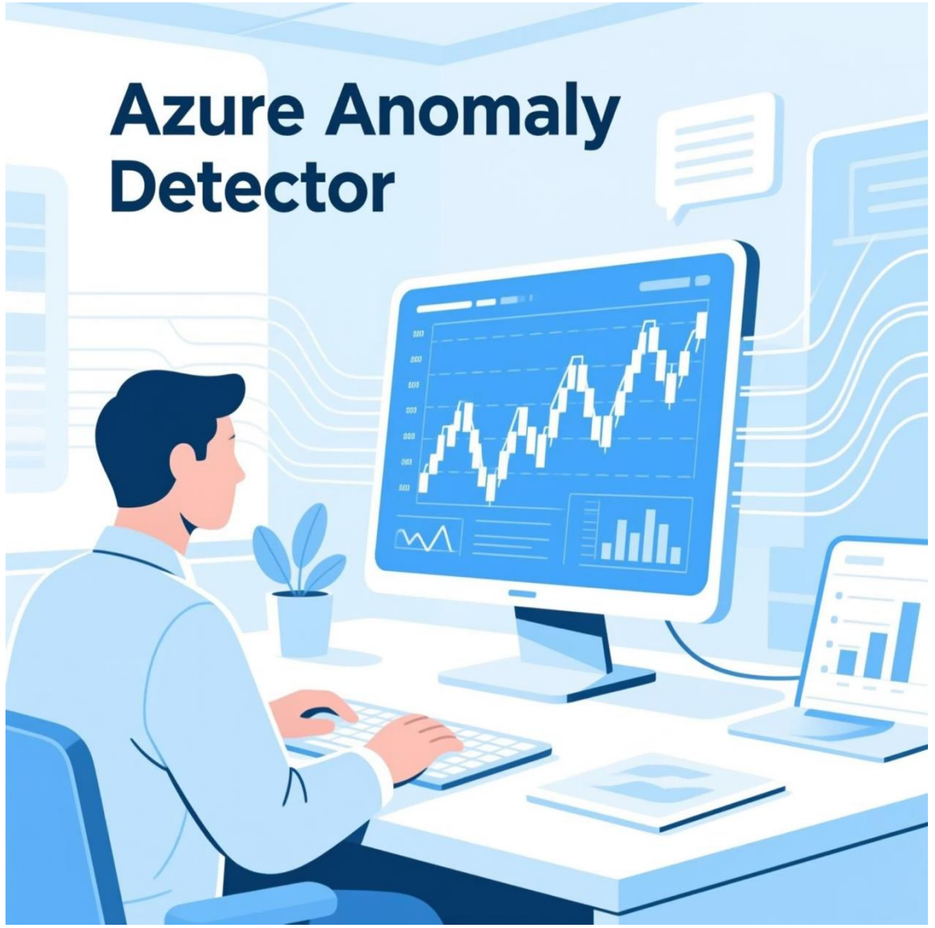
Alternativas en GCP y Azure (Detección de Fraude)

Google Cloud Platform (GCP)

- **Ingesta/Procesamiento:** Pub/Sub + Dataflow
- **ML:** Vertex AI Prediction
- **Característica Destacada:** Vertex AI Feature Store para gestionar gestionar características en tiempo real

Microsoft Azure

- **Ingesta/Procesamiento:** Event Hubs + Stream Analytics o Azure Functions
- **ML:** Azure ML Endpoints
- **Servicio Gestionado:** Azure AI Anomaly Detector (más genérico, pero aplicable)



Transición: De los Casos de Uso a los los Desafíos Reales

Hemos visto cuatro arquitecturas de referencia.

Parece que "sólo" hay que conectar servicios.

Pero... ¿cuáles son los problemas reales que encontraremos en el camino?

A continuación: Los 7 desafíos de implementar IA en el cloud.

Desafíos Comunes (Parte 1: Coste, Complejidad, Escalabilidad) Escalabilidad)

1

Gestión de Costes

- **GPUs:** Extremadamente caras. Optimizar su su uso es crucial.
- **Transferencia de Datos (Egress):** Mover datos fuera de la nube puede generar facturas inesperadas.

2

Complejidad de MLOps

Automatizar el ciclo de vida (CI/CD/CT/CM) es difícil. Requiere herramientas específicas y una cultura organizativa adecuada.

3

Escalabilidad y Rendimiento en Inferencia

Garantizar baja latencia para millones de usuarios usuarios es un reto de ingeniería (autoescalado, (autoescalado, optimización de modelos).

Desafíos Comunes (Parte 2: Seguridad, Datos, Talento) Talento)

1

Seguridad y Cumplimiento Normativo

Proteger datos sensibles (como vimos en el caso médico) no es opcional opcional y requiere un diseño cuidadoso desde el principio.

2

Vendor Lock-in y Portabilidad

Usar servicios muy específicos de un proveedor (ej. Amazon Personalize) Personalize) dificulta la migración a otra nube.

3

Gestión de Datos y Feature Engineering

"Garbage in, garbage out". La calidad de los datos y las características características determina el éxito del modelo.

4

Disponibilidad de Talento Especializado

Los perfiles que dominan tanto IA como Cloud son escasos y muy muy demandados en el mercado laboral.

Mirando al Futuro: Tendencias en Cloud para IA

El panorama de Cloud e IA evoluciona a una velocidad vertiginosa.

Vamos a explorar las 7 tendencias que están definiendo la próxima generación de infraestructuras de IA.

Tendencia 1: IA Generativa en el Cloud

¿Qué es?

Modelos capaces de crear nuevo contenido (texto, imágenes, código). código).

Servicios Cloud

- **AWS:** Amazon Bedrock (acceso a múltiples modelos fundacionales como API)
- **GCP:** Vertex AI Model Garden (acceso a modelos de Google y otros) otros)
- **Azure:** Azure OpenAI Service (acceso a modelos de OpenAI como GPT-4)

Implicaciones

Infraestructura optimizada para entrenamiento y fine-tuning de LLMs, LLMs, nuevas consideraciones de coste y ética.

Tendencia 2: Edge AI y Cloud-to-Edge

¿Qué es?

Ejecutar modelos de IA directamente en dispositivos (móviles, coches, sensores sensores IoT) en lugar de en la nube.

¿Por qué?

- Baja latencia
- Privacidad de datos
- Funcionamiento offline

Servicios Cloud para gestionar el Edge

AWS

IoT Greengrass, SageMaker Edge Manager

Azure

IoT Edge

GCP

Vertex AI Edge

Tendencia 3: Serverless AI

¿Qué es?

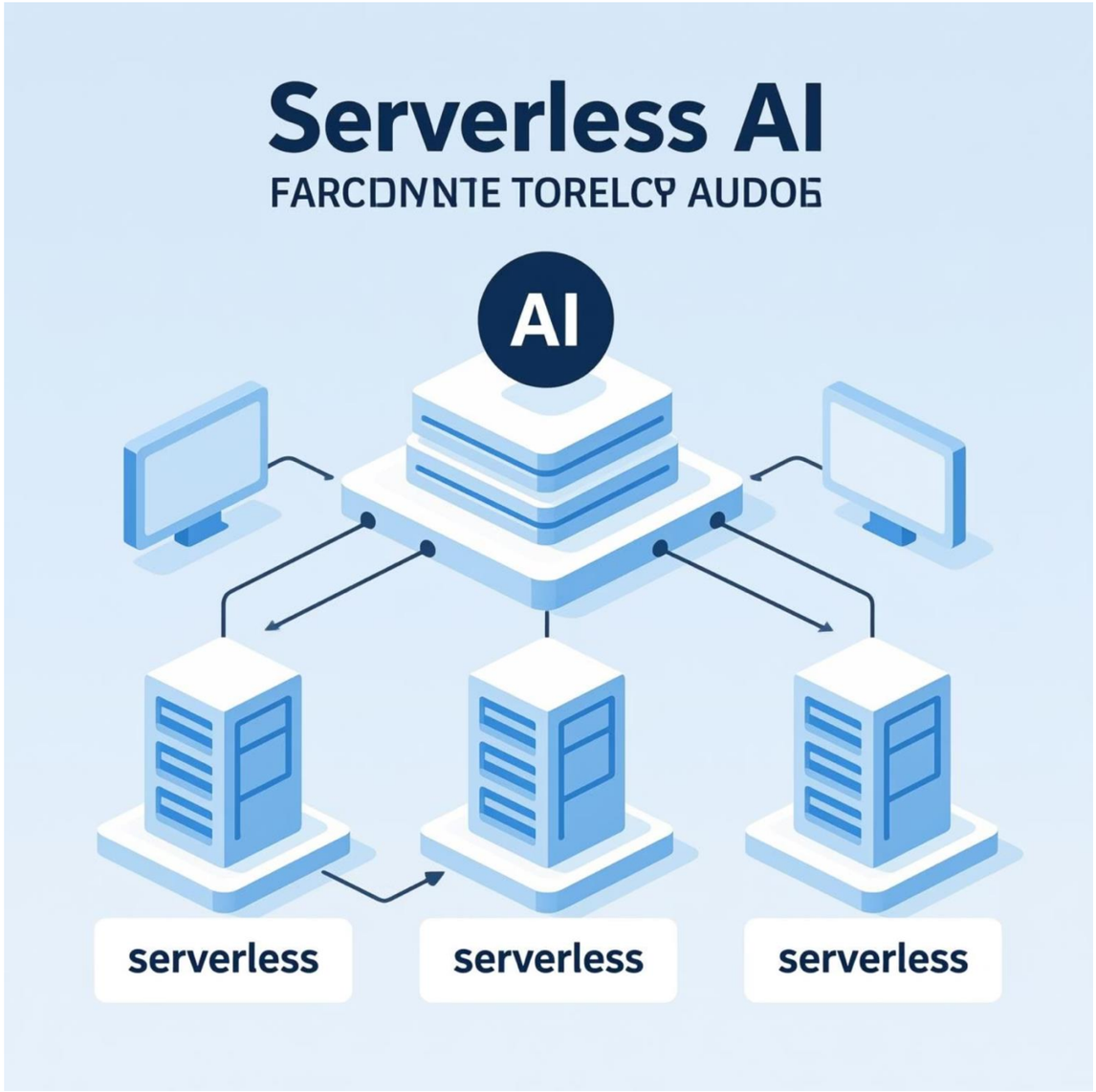
Usar arquitecturas sin servidor para componentes de IA.

Componentes

- **FaaS (Functions as a Service):** Lambda, Cloud Functions, Azure Functions para para preprocesamiento o inferencia de modelos ligeros
- **Contenedores Serverless:** Fargate, Cloud Run para desplegar APIs de modelos sin gestionar servidores

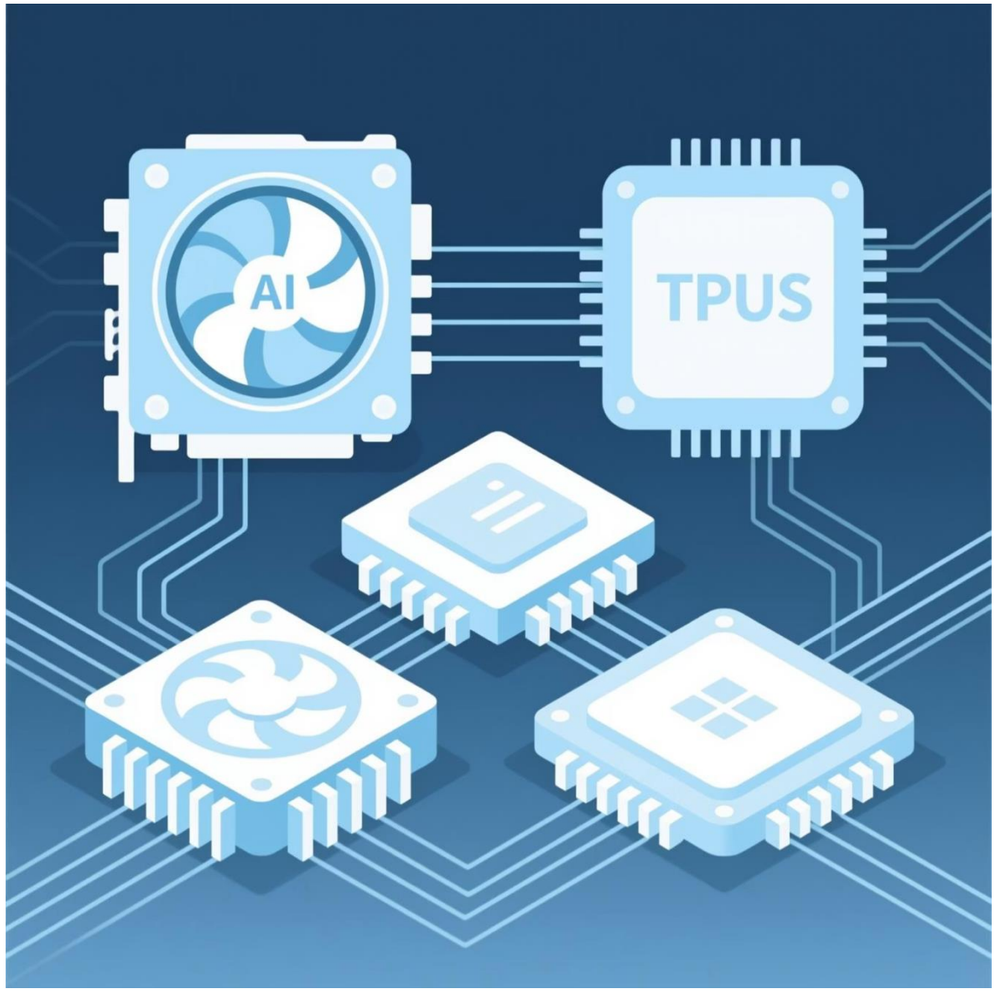
Ventajas

- Pagar solo por el uso real
- Escalado automático a cero
- Menor carga operativa



Tendencia 4: Optimización de Hardware para IA

La carrera por el hardware más eficiente para IA está en pleno apogeo.



Tipos de Aceleradores

GPUs

Siguen siendo el estándar (NVIDIA)

Chips para Entrenamiento

Google TPUs (Tensor Processing Units)

AWS Trainium

Chips para Inferencia

AWS Inferentia

Tendencia 5: IA Responsable y Ética en la Nube

¿Qué es?

La necesidad de construir sistemas de IA que sean justos, transparentes y explicables.

Herramientas Cloud

- **AWS:** SageMaker Clarify (detecta sesgos en datos y modelos, explica predicciones)
- **GCP:** Vertex AI Explainable
  Explainable AI
- **Azure:** Azure ML Responsible AI
  Dashboard

Conceptos Clave

Explicabilidad (XAI)

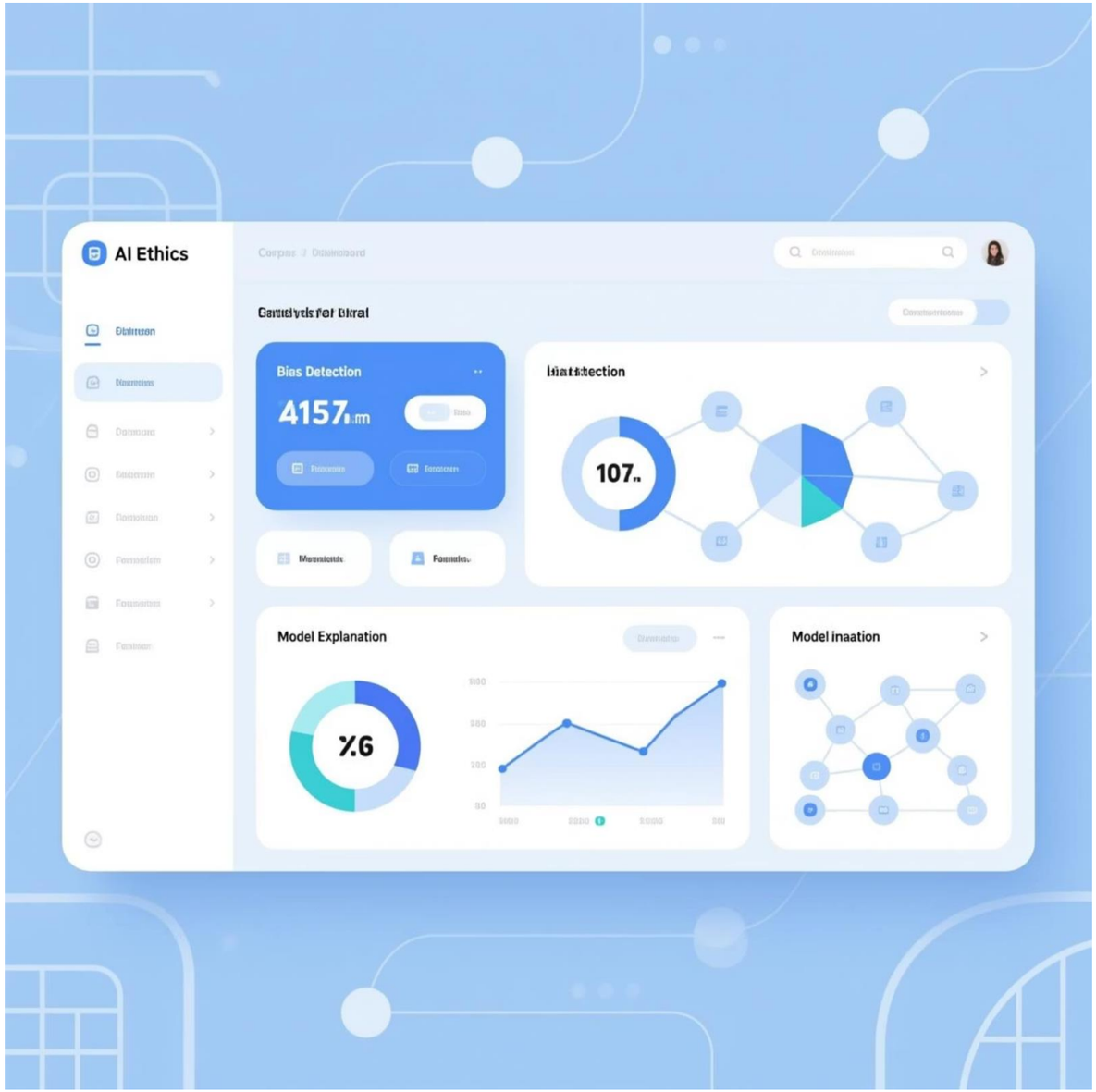
Entender por qué un modelo toma ciertas decisiones

Equidad (fairness)

Garantizar que el modelo no discrimina a grupos específicos

Mitigación de sesgos

Identificar y corregir sesgos en datos y modelos



Tendencia 6: Data-Centric AI

El Foco Cambia

Del "model-centric" (mejorar el algoritmo) al "data-centric" (mejorar los datos). los datos).

La Realidad

La calidad y cantidad de los datos es, a menudo, más importante que la última la última arquitectura de modelo.

Herramientas Cloud que lo soportan

- Servicios de etiquetado de datos (SageMaker Ground Truth)
- Feature Stores
- Herramientas de gestión y versionado de datos

"En lugar de pasar meses ajustando el modelo, dedica ese tiempo a mejorar la calidad de tus datos de entrenamiento: limpia etiquetas, añade ejemplos de casos difíciles, etc."

Tendencia 7: Quantum Computing (Visión a Largo Plazo)

¿Qué es?

Un paradigma de computación completamente nuevo basado en la mecánica cuántica.

Potencial Impacto en IA

Resolver problemas de optimización y machine learning que son intratables para los ordenadores clásicos.

Servicios Cloud (Experimentales)

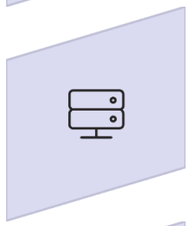
- Amazon Braket
- Azure Quantum

Estado Actual: Todavía en fase de investigación y desarrollo, no para producción.

Repaso General del Curso

Fundamentos

¿Qué es el Cloud? (IaaS, PaaS, SaaS)

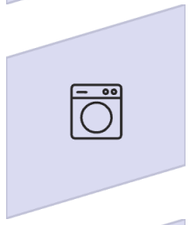


Core Services

Computación (EC2, Lambda), Almacenamiento (S3), Bases de Datos (RDS, DynamoDB), Redes (VPC)

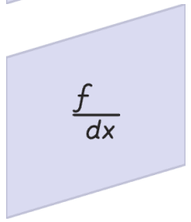
Ecosistema de Datos

Ingesta (Kinesis), Procesamiento (Glue), Almacenamiento (S3), Análisis (Athena)



Plataforma de ML

Desde la preparación de datos hasta el despliegue y monitorización (El ciclo de vida de SageMaker)



Integración

Cómo todos estos servicios se combinan para crear soluciones de IA end-to-end

¿Y ahora qué? Próximos Pasos

Practicar

¡La teoría no es suficiente! Usad las cuentas de AWS Educate / Azure for for Students / GCP Free Tier.

- Replicad alguna de las arquitecturas que hemos visto hoy a pequeña escala

- Cread proyectos personales para aplicar lo aprendido

Certificaciones

- **Fundacionales:** AWS Cloud Practitioner, Azure AZ-900, GCP Cloud Digital Leader
- **Especializadas:** AWS Certified Machine Learning - Specialty, Azure DP-100, GCP Professional Machine Learning Engineer

Mantenerse al día

Discusión Abierta y Preguntas Finales

Preguntas sobre los casos de estudio

¿Cómo aplicarías estas arquitecturas a un problema específico?

Preguntas sobre las tendencias futuras

¿Qué tendencia consideras más relevante para tu carrera?

Preguntas sobre el curso en general

Dudas sobre conceptos, servicios o implementaciones

Cualquier otra duda

Ahora es el momento de resolver cualquier inquietud

¡Gracias!