Cloud Computing para Inteligencia Artificial

Sesión 9 - Plataformas de ML Gestionadas II: Despliegue, Monitorización y Comparativa de Plataformas

Agenda de la Sesión

01

Recapitulación Rápida

Del entrenamiento a la producción: dónde estamos en el ciclo de vida ciclo de vida de ML.

02

Fase de Despliegue con SageMaker

De los artefactos a la inferencia en tiempo real y batch.

03

Monitorización y Mantenimiento

Asegurando la calidad del modelo en producción.

04

Automatización y MLOps

SageMaker Pipelines como orquestador de flujos de trabajo.

05

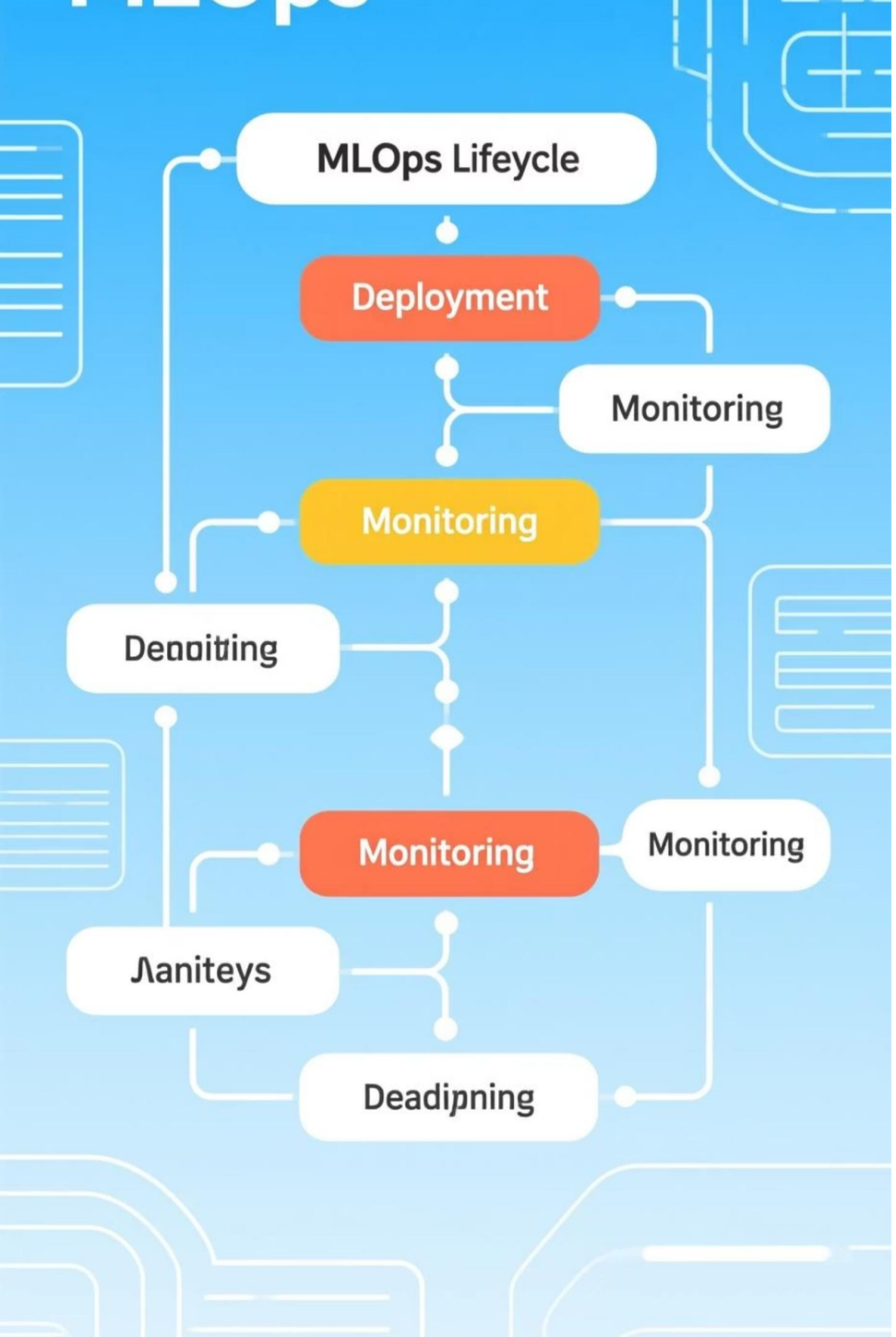
El Ecosistema Cloud

Comparativa detallada con GCP Vertex AI y Azure ML.

06

Resumen y Próximos Pasos

Consolidación de conceptos y actividades futuras.



El Ciclo de Vida del ML: ¿Dónde Estamos?

Tras haber explorado las fases iniciales del ciclo de ML (comprensión del negocio, adquisición y preparación de datos, entrenamiento y evaluación), hoy nos adentramos en lo que se conoce como la "última milla" del ML:

Despliegue

Cómo llevar los modelos entrenados a producción de forma eficiente

Monitorización

Vigilancia continua del rendimiento y deriva de los modelos

Mantenimiento

Actualización y mejora de modelos a lo largo del tiempo

El objetivo final: entregar valor real y sostenido a través de modelos de IA que operan de forma de forma fiable en el mundo real.

Artefactos del Modelo: El "Paquete" de Nuestro Trabajo

Tras un job de entrenamiento exitoso en SageMaker, obtenemos un conjunto de **artefactos artefactos** que se almacenan en un bucket de S3 como un archivo comprimido:

model.tar.gz

Este "paquete" es la unidad fundamental para el despliegue, siendo autocontenido y autocontenido y versionable.

Archivos del modelo

Pesos y arquitectura (.pth, .pb, .pkl)

Código de inferencia

inference.py con funciones model_fn(), input_fn(), predict_fn(), output_fn()

Requerimientos

Dependencias en requirements.txt

Inferencia en Tiempo Real: SageMaker Real-time Endpoints

Concepto

Un servicio HTTP/S persistente que aloja tu aloja tu modelo para predicciones de baja baja latencia. Ideal para:

- Chatbots interactivos
- Sistemas de recomendación en tiempo tiempo real
- Detección de fraude en transacciones

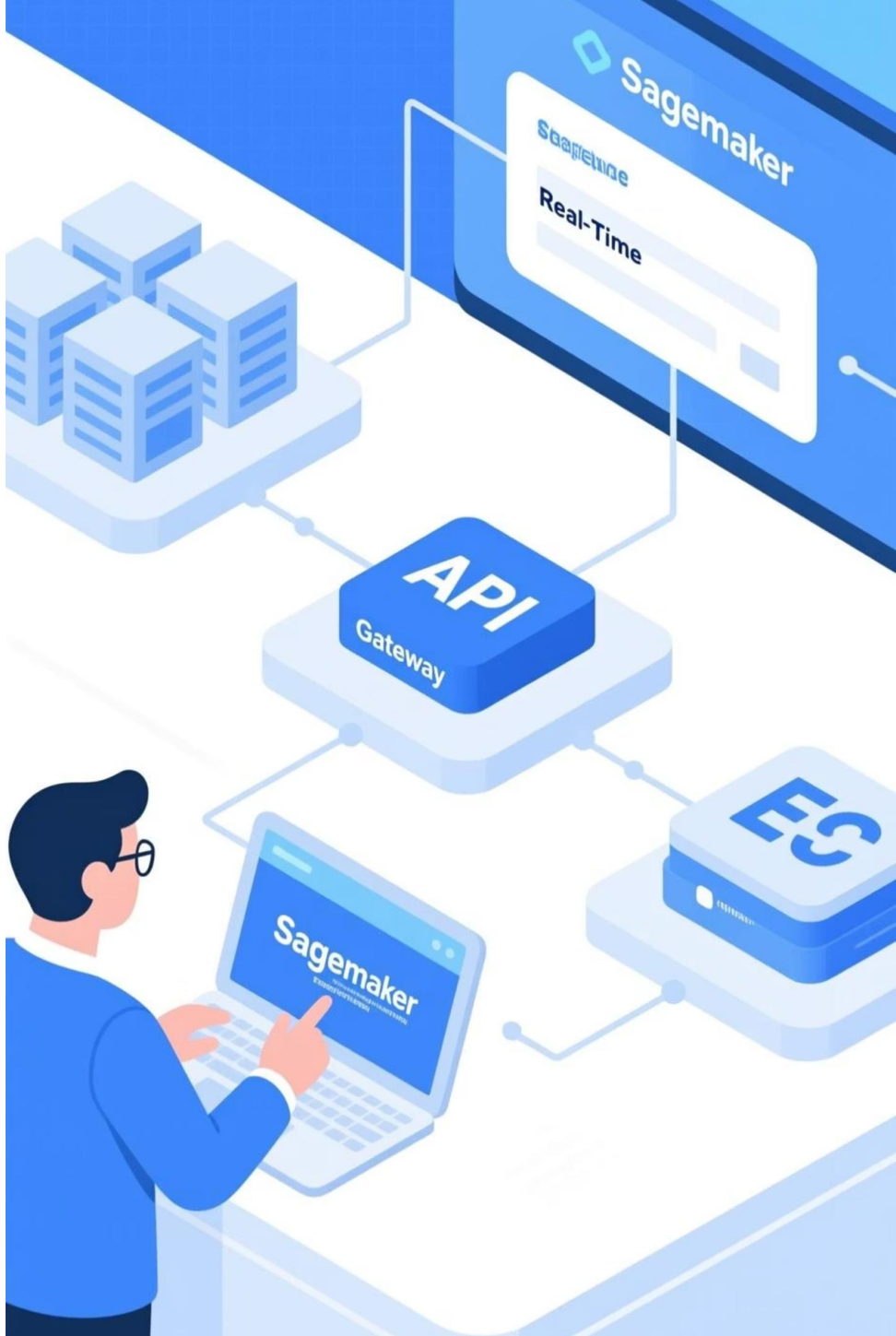
Características Clave

Auto Scaling

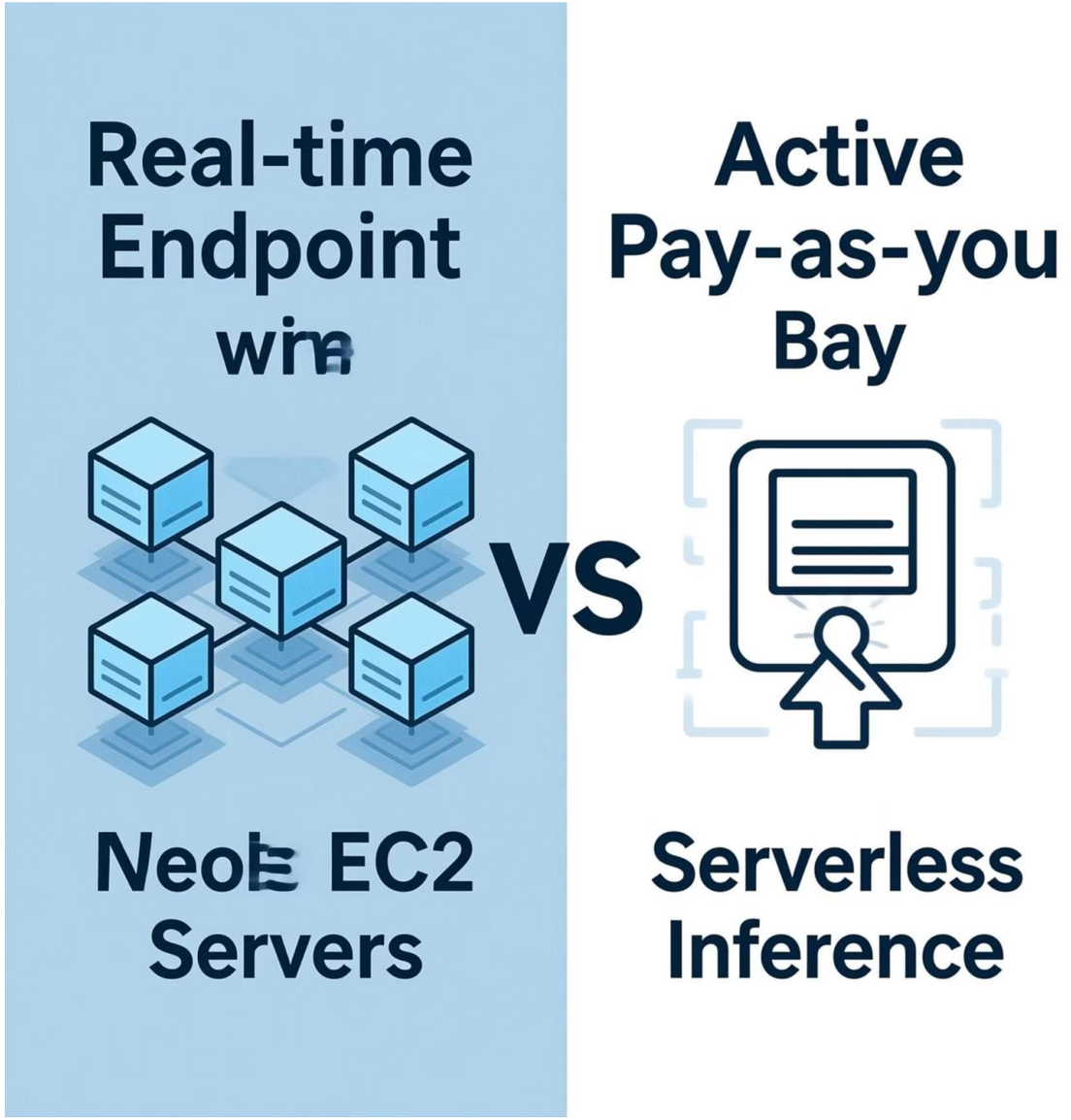
Ajusta automáticamente el número de número de instancias EC2 según la carga

Variantes de Modelos

Permite desplegar múltiples modelos en el mismo endpoint para A/B testing



Inferencia sin Servidor: SageMaker Serverless Inference



Concepto

Opción de inferencia que abstrae completamente la gestión de servidores. No hay instancias EC2 que configurar.

Caso de Uso Ideal

Cargas de trabajo **intermitentes o esporádicas** con largos periodos de de inactividad.

Ventajas

- Pagas solo por el tiempo de cómputo y datos procesados
- Sin coste si no se usa
- Escalado automático de cero a lo necesario

Consideración

Posible "arranque en frío" (cold start) con mayor latencia inicial

Inferencia Asíncrona: Para Tareas Pesadas

Diseñada para predicciones que no requieren respuesta inmediata. El cliente envía una petición y recibe confirmación de aceptación, mientras el procesamiento ocurre en segundo plano.

Caso de Uso Ideal

- Payloads grandes (hasta 1GB): imágenes de alta resolución, documentos extensos
- Tiempos de procesamiento largos: modelos complejos que tardan minutos

Flujo de Trabajo

1. Cliente envía petición al endpoint
2. Petición encolada en SQS gestionado por SageMaker
3. SageMaker procesa la petición de la cola
4. Resultado depositado en S3
5. Notificación opcional vía SNS

Transformación por Lotes: SageMaker Batch Transform

Concepto

Permite realizar inferencias sobre un conjunto de datos completo (batch) de una (batch) de una sola vez. No requiere un endpoint persistente, lo que lo hace lo hace muy rentable.

Casos de Uso Ideales

- Generación de informes diarios de segmentación
- Puntuación de leads de marketing al final del día
- Pre-procesamiento de grandes volúmenes de datos

Flujo de Trabajo

Inicias un "trabajo de transformación"

SageMaker provisiona la infraestructura necesaria

Lee datos de entrada desde S3

Ejecuta predicciones sobre todo el dataset

Guarda resultados en otro bucket S3

Infraestructura se desprovisiona automáticamente

Despliegue en el Borde: SageMaker Edge Manager & Neo

No todas las inferencias ocurren en la nube. A veces, se necesita IA en dispositivos con:

- Conectividad limitada
- Requisitos de latencia ultra-baja
- Procesamiento local por privacidad

SageMaker Neo

Compilador de modelos que optimiza para hardware específico (NVIDIA, Intel, ARM, etc.).

Genera ejecutables hasta 25 veces más rápidos y más pequeños sin pérdida de precisión.

SageMaker Edge Manager

Servicio para operar, monitorizar y actualizar flotas de modelos en dispositivos edge.

Empaqueta el modelo optimizado con un agente de software y gestiona el despliegue seguro.

La Realidad de la Producción: ¿Por Qué Monitorizar?

Un modelo desplegado no es el final del camino. El mundo real cambia constantemente, y el rendimiento del modelo puede degradarse con el tiempo.

1

Data Drift

La distribución estadística de los datos de entrada cambia. Ejemplo: un modelo de recomendación de moda entrenado con datos de verano empieza a recibir datos de invierno.

2

Concept Drift

La relación entre variables de entrada y objetivo cambia. Ejemplo: en un modelo de fraude, los estafadores inventan nuevas técnicas que el modelo nunca ha visto.

3

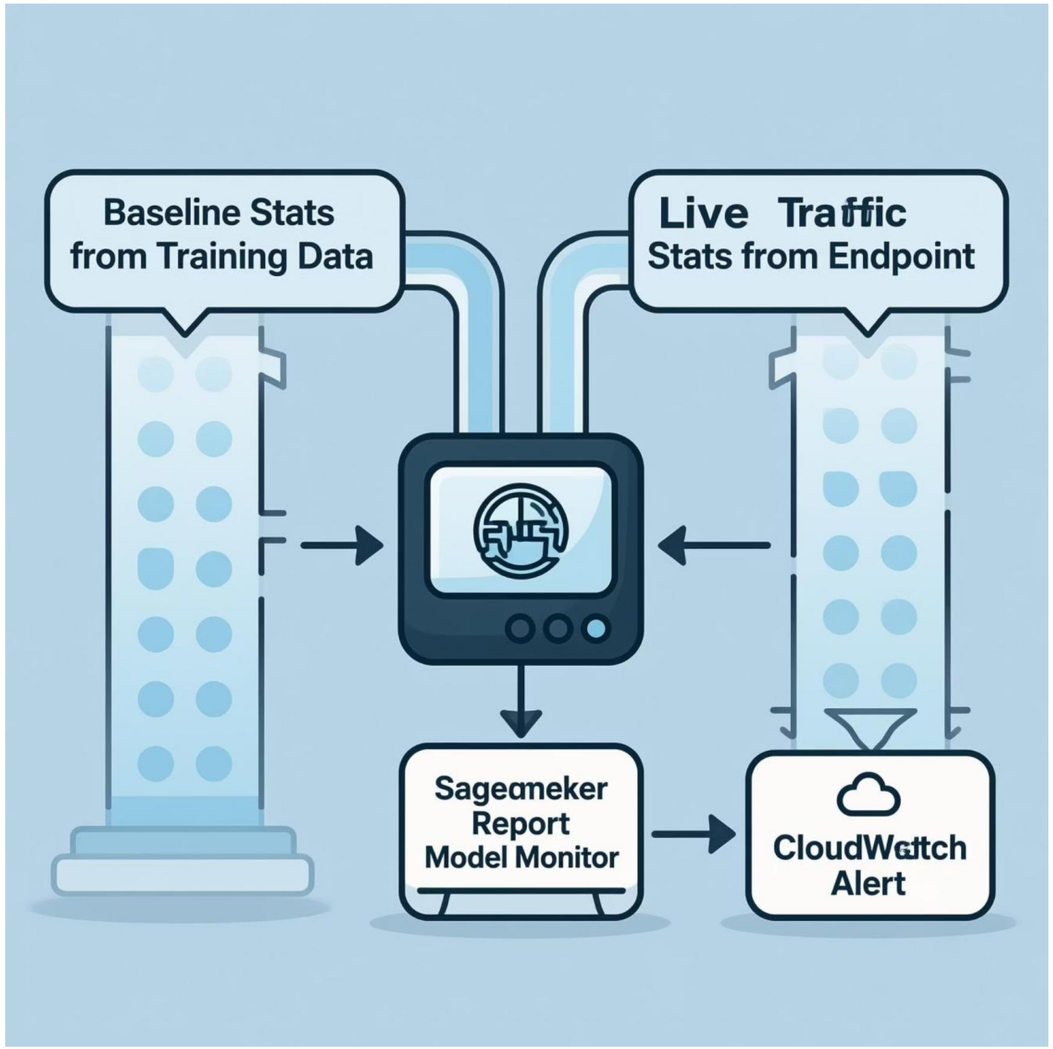
Bias Drift

El modelo empieza a hacer predicciones injustas o sesgadas para para ciertos subgrupos demográficos, demográficos, afectando a la equidad equidad del sistema.

SageMaker Model Monitor

Función Principal

Detecta automáticamente la desviación de la calidad de un modelo en producción.



¿Cómo funciona?

Línea Base (Baseline)

Durante el entrenamiento, se crea una línea base de estadísticas y restricciones de los restricciones de los datos

Captura de Datos

El endpoint captura un porcentaje del tráfico de inferencia en tiempo real

Comparación

Periódicamente, un trabajo compara las estadísticas de los datos capturados con capturados con la línea base

Alerta

Si se detectan violaciones, se genera un informe y se puede lanzar una alerta una alerta en CloudWatch

Explicabilidad y Sesgo: SageMaker Clarify

Función Principal

Proporciona visibilidad sobre el comportamiento de los modelos y posibles sesgos. Componente esencial para la IA Responsible.

Análisis de Sesgo (Pre-entrenamiento)

Mide sesgos en el conjunto de datos inicial

Análisis de Sesgo (Post-entrenamiento)

Evalúa si el modelo produce predicciones sesgadas

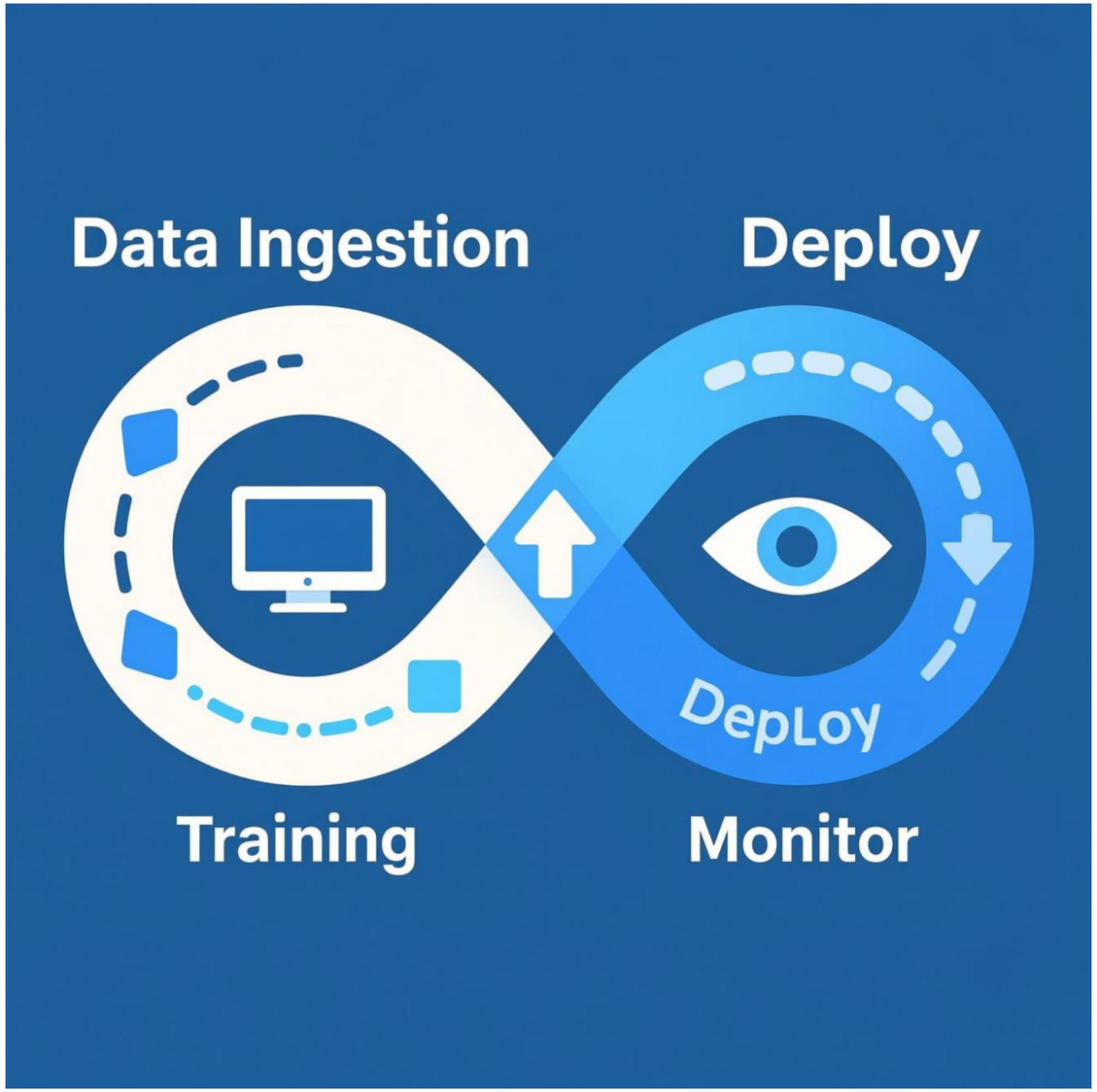
Equidad (Fairness) y Ética. Se centra en la salida de las predicciones en relación con las características protegidas (ej., género, etnia, edad).

Métricas de Disparidad (ej., Tasa de Falsos Positivos Dispar, que compara la tasa entre un grupo minoritario y el grupo de referencia).

Explicabilidad (Explainability)

- **Transparencia y auditabilidad**. Ayuda a entender por qué un modelo tomó una decisión específica
- Utiliza algoritmos como SHAP para calcular la importancia de cada feature
- Genera informes visuales cruciales para reguladores, stakeholders y depuración

MLOps: El DevOps para Machine Learning



MLOps (Machine Learning Operations) es una cultura y conjunto de prácticas que buscan unificar el desarrollo de modelos de ML (Dev) con su operación en producción (Ops).

Objetivos

- Automatizar el ciclo de vida completo del ML
- Aumentar la velocidad de entrega de modelos
- Mejorar la fiabilidad y reproducibilidad
- Facilitar la colaboración entre equipos

En lugar de ejecutar notebooks manualmente, creamos flujos de trabajo automatizados, versionados y repetibles.

Orquestación con SageMaker Pipelines

El primer servicio de CI/CD (Integración Continua / Despliegue Continuo) creado específicamente para Machine Learning, permitiendo definir todo el flujo de trabajo como un Grafo Acíclico Dirigido (DAG).

ProcessingStep

Para la preparación de datos y transformación de features

TrainingStep

Para entrenar el modelo usando diferentes algoritmos

EvaluationStep

Para evaluar el rendimiento con datos de test

ConditionStep

Para tomar decisiones basadas en métricas

RegisterModelStep

Para registrar nuevas versiones en el registro

ModelStep

Desplegarlo en producción

A veces se requiere paso manual (otro equipo)

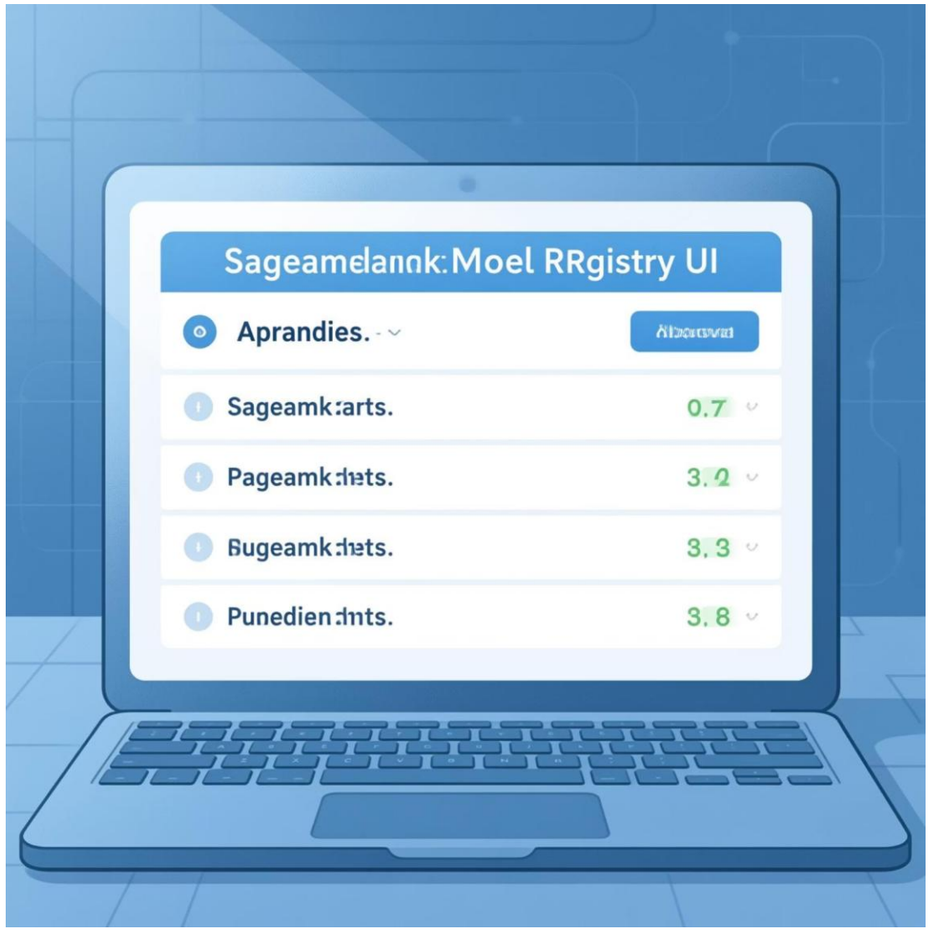
Normalmente acompañado por un ConditionStep

Gobernanza de Modelos: SageMaker Model Registry

¿Qué es?

Un catálogo centralizado para gestionar, versionar y seguir la trazabilidad d
modelos de ML.

Funciona como un "almacén de artefactos" para modelos aprobados, facilitando su gestión a lo largo del tiempo.



Funcionalidades Clave

Versionado

Cada modelo re-entrenado se registra como nueva versión dentro de un "Grupo de Modelos"

Metadatos

Almacena métricas de evaluación, ubicación de artefactos en S3 y parámetros de entrenamiento

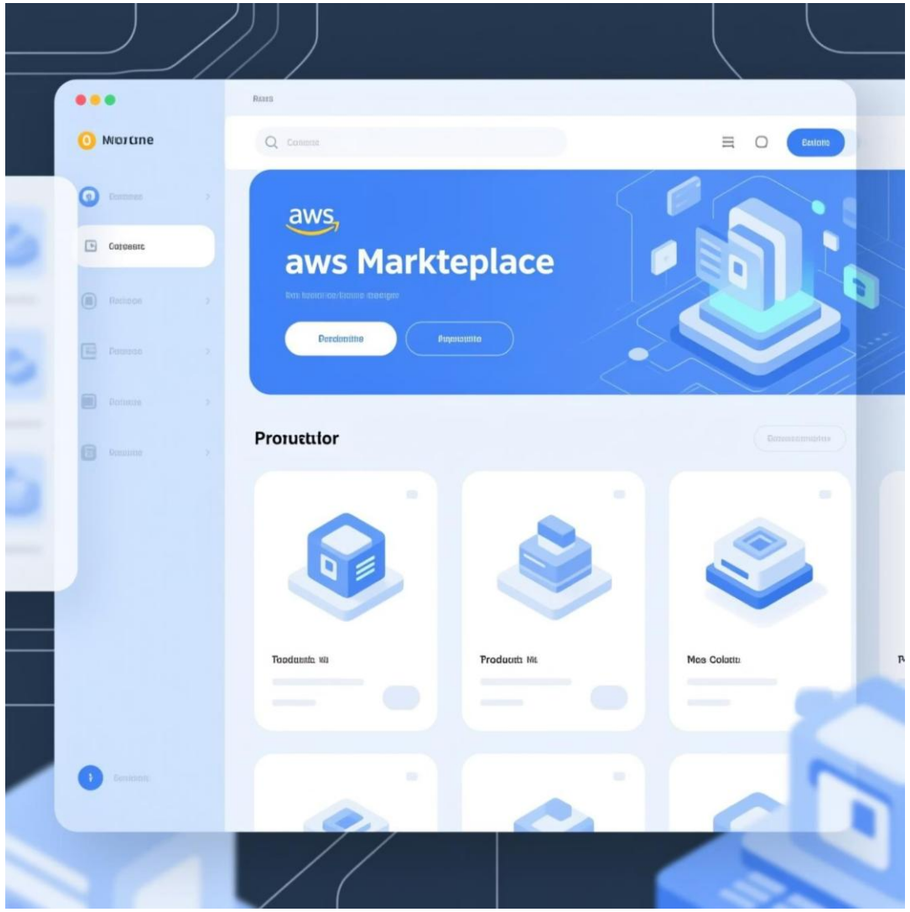
Estado de Aprobación

Gestiona el ciclo de vida: Pending, Approved, Rejected. Los pipelines CI/CD solo solo despliegan modelos "Approved"

Marketplace de Algoritmos y Modelos

Concepto

Un "App Store" para Machine Learning. SageMaker Marketplace proporciona proporciona acceso a cientos de algoritmos y modelos pre-entrenados de entrenados de terceros.



¿Qué puedes encontrar?

Algoritmos

Soluciones para problemas específicos (ej: análisis geoespacial, modelos financieros)

Modelos Pre-entrenados

Listos para usar en tareas como reconocimiento de objetos o análisis de sentimientos

Ventajas

- Acelera el desarrollo al no partir desde cero
- Permite probar soluciones de pago antes de comprarlas
- Se despliegan igual que tus propios modelos

¡Momento Interactivo! Elige el Despliegue Correcto

Para los siguientes casos de uso, ¿qué método de despliegue de SageMaker sería el más adecuado y por qué?

Escenario 1

Un servicio de traducción de texto en una una app de chat que necesita respuestas respuestas instantáneas y tiene un flujo de flujo de usuarios constante.

Escenario 2

Una firma de arquitectura necesita procesar 500 modelos 3D de alta resolución cada noche para detectar fallos estructurales. Cada modelo tarda 15 minutos en analizarse.

Escenario 3

Una startup quiere añadir una función de "etiquetado de fotos" a su app, pero esperan muy poco tráfico durante los primeros 6 meses.

El Ecosistema de Plataformas de ML en ML en la Nube

AWS SageMaker es un líder del mercado, pero Google Cloud Platform (GCP) y Microsoft Azure ofrecen plataformas muy competitivas y potentes.

La elección depende de:

- El ecosistema cloud existente en la empresa
- Las herramientas de datos y análisis con las que se necesita integración
- La experiencia previa del equipo
- Requisitos específicos del proyecto

Aunque cada plataforma tiene sus fortalezas, todas comparten el objetivo común de simplificar el desarrollo y despliegue de modelos de ML en producción.

A continuación analizaremos en detalle las detalle las dos alternativas principales a a SageMaker:

- GCP Vertex AI
- Azure Machine Learning

Google Cloud Platform - Vertex AI

Filosofía

Una plataforma unificada que busca integrar todas las herramientas de ML en una sola API y UI.

Componentes Clave y sus análogos en SageMaker

Plataforma Unificada

Similar a SageMaker Studio

Vertex AI Workbench

Entornos Jupyter gestionados, como los Notebook Instances

Vertex AI Training & Prediction

Equivalente a Training Jobs y Endpoints

Vertex AI Pipelines

Orquestador basado en Kubeflow Pipelines / TFX

Fortalezas y Diferenciadores

- **Integración con BigQuery:** Entrenar modelos directamente en BigQuery con SQL (BigQuery ML)
- **Infraestructura de IA de Google:** Acceso fácil a TPUs (Tensor Processing Units)
- **AutoML:** Capacidades muy potentes y maduras, fortaleza histórica de GCP
- **Vertex AI Matching Engine:** Servicio de vanguardia para búsqueda de similitud de vectores

Microsoft Azure - Azure Machine Learning

Filosofía

Fuerte enfoque en la accesibilidad para diferentes perfiles y en la IA Responsable.

Componentes Clave

- **Azure ML Workspace:** Entorno colaborativo central
- **Azure ML Studio:** UI web completa que incluye:
  - Notebooks para desarrollo basado en código
  - Automated ML (AutoML) para entrenamiento automático
  - Designer: interfaz visual drag-and-drop
- **Endpoints (Online y Batch):** Análogos a SageMaker
- **MLOps:** Integración nativa con Azure DevOps

Fortalezas y Diferenciadores

Herramientas Visuales

Designer facilita la entrada a personas sin profundo conocimiento de programación

Tabla Comparativa: SageMaker vs. Vertex AI vs. Azure ML

Característica	AWS SageMaker	GCP Vertex AI	Azure Machine Learning
Facilidad de Uso	Muy completo, pero con curva de aprendizaje más pronunciada debido a la a la granularidad de sus servicios	UI unificada y limpia. Integración con BigQuery ML es un plus para analistas	Muy accesible, especialmente con Designer Designer visual y AutoML. Ideal para equipos mixtos
Integración Ecosistema	Excelente con stack AWS (S3, Lambda, CloudWatch)	Inmejorable con ecosistema de datos GCP (BigQuery, GCS, Looker)	Profunda con mundo Microsoft (Azure DevOps, Power BI, GitHub)
Flexibilidad (BYOM)	Muy alta. Soporte nativo para contenedores personalizados (BYOC), scripts, etc.	Alta. Fuerte soporte para contenedores personalizados y código abierto (Kubeflow)	Alta. Soporte para contenedores, pero con enfoque más guiado
MLOps y Gobernanza	Madura con Pipelines, Model Registry, Clarify y Model Monitor	Potente con Pipelines. Fuerte en linaje de datos y artefactos	Excelente integración en Azure DevOps y dashboard de IA Responsable líder
Coste	Modelo granular, pagas por cada componente. Puede ser complejo de predecir	Precios competitivos, especialmente en AutoML	Modelos flexibles, incluyendo niveles "enterprise"
Hardware Especial	GPUs, Inferentia, Trainium	GPUs y TPUs	GPUs y FPGAs

Conclusiones de la Comparativa

No hay un "ganador" absoluto. La mejor plataforma depende del específico de tu organización.

Elige AWS SageMaker si:

- Tu empresa ya está en AWS
- Necesitas máxima granularidad y control
- Quieres acceder a su amplio ecosistema de servicios especializados

Elige GCP Vertex AI si:

- Tu estrategia de datos gira en torno a BigQuery
- Valoras una UI muy unificada
- Necesitas el rendimiento de las TPUs TPUs para deep learning a gran escala escala

Elige Azure Machine Learning si:

- Tu organización utiliza el stack de Microsoft/Azure DevOps
- Tienes un equipo con diversos niveles técnicos (gracias al Designer)
- La IA Responsable es una prioridad estratégica

Resumen de la Sesión

Hemos cubierto el ciclo de vida de un modelo después del entrenamiento, explorando las fases cruciales para llevar modelos a producción.

1

Despliegue en SageMaker

Aprendimos las opciones para servir modelos: inferencia en tiempo real (Real-time, Serverless), Serverless), asíncrona y por lotes (Batch Transform)

2

Monitorización

Comprendimos la importancia de vigilar el data drift con Model Monitor y garantizar la equidad con Clarify

3

MLOps

Vimos cómo SageMaker Pipelines automatiza y orquesta todo el flujo de trabajo de ML para lograr para lograr CI/CD efectivo

4

Ecosistema Cloud

Comparamos las fortalezas y enfoques de SageMaker, Vertex AI y Azure ML, entendiendo que la entendiendo que la elección depende del contexto

¿Preguntas?