Preguntas Clave de la Sesión 8

Bloque 1: El Ciclo MLOps y Desafíos Clave

Pregunta de Diferencia de Conceptos: Drift

Dentro de la fase de Monitorización, ¿cuál es la diferencia crítica entre Data Drift y Concept Drift, y qué implicación tiene un caso como el *Concept Drift* para el modelo en producción?

Pregunta de MLOps DIY: El Desafío más Costoso

Una organización debe elegir entre resolver los desafíos de Escalabilidad o Reproducibilidad en un enfoque DIY. ¿Cuál de estos desafíos principales del MLOps "Do It Yourself" se considera el más costoso de resolver en términos de infraestructura y gestión de cómputo, y por qué?

Pregunta de Gobernanza y Sesgo

¿Por qué la medición y el Análisis de Sesgos (*Bias*) son esenciales para la Gobernanza del modelo, y qué herramienta de SageMaker proporciona esta funcionalidad de equidad (*fairness*)?

Bloque 2: Optimización y Orquestación de Plataformas

Pregunta de Estrategia de Costes

Para un trabajo de entrenamiento de Deep Learning largo y costoso, ¿qué estrategia de SageMaker permite reducir los costes hasta en un 90% y cómo gestiona la plataforma la interrupción de la capacidad?

Pregunta de MLOps y Abstracción

Las plataformas gestionadas ofrecen la Democratización del MLOps. ¿Cuál es la principal ventaja que aporta un servicio como SageMaker Pipelines para la colaboración entre equipos, y qué tipo de estructura utiliza para automatizar el flujo?

Pregunta de Comparativa Cloud y Ecosistema

Tu empresa está considerando migrar su *stack* de datos a GCP, ya que su estrategia de datos gira en torno a BigQuery. ¿Qué plataforma de ML Cloud sería la opción estratégica más natural para integrar el entrenamiento de modelos directamente desde esa fuente de datos?

Escenario A: Despliegue de Alto Tráfico y Baja Latencia

Descripción: Eres una gran plataforma de *e-commerce* y necesitas un modelo de recomendación que responda en **milisegundos** a cientos de miles de **peticiones constantes** por hora, 24/7.

- **Opciones de Despliegue:** 1. Inferencia Asíncrona, 2. Real-time Endpoint (Aprovis.), 3. Batch Transform.

Escenario B: MLOps y Despliegue Condicional

Descripción: Has entrenado un nuevo modelo y solo debe pasar a despliegue automático si su **F1-score es superior al 85%**. Si no lo es, el *pipeline* debe detenerse[3].

- **Opciones de Orquestación:** 1. ProcessingStep, 2. ConditionStep, 3. Model Monitor.

Escenario C: Eficiencia de Costes e Intermitencia

Descripción: Estás desarrollando un modelo interno para el departamento de RRHH que predice la rotación de personal, pero solo se ejecuta de forma **esporádica o intermitente** (una o dos veces al mes).

- **Opciones de Inferencia:** 1. Real-time Endpoint (Aprovis.), 2. Serverless Inference, 3. Batch Transform.

Escenario D: Detección de Drift en Producción

Descripción: Tu modelo de clasificación de spam ha sido desplegado, pero las nuevas campañas de *phishing* han cambiado la distribución de los correos de entrada (**Data Drift**). ¿Qué servicio **compara automáticamente** el tráfico en vivo con la línea base de entrenamiento para generar una alerta?

- **Opciones de Monitorización:** 1. SageMaker Debugger, 2. SageMaker Model Monitor, 3. CloudWatch Logs.

Escenario A: Contenedores y Algoritmos

Pregunta sobre Opciones de Entrenamiento

Tienes un modelo para clasificar comentarios de clientes utilizando un algoritmo de procesamiento de lenguaje natural llamado **BlazingText**. Quieres entrenarlo a gran escala con SageMaker, pero no tienes tiempo para escribir código de entrenamiento ni configurar un contenedor. ¿Cuál de las opciones de entrenamiento de SageMaker es la más adecuada para ti y por qué?

Pregunta sobre Contenedores y Flexibilidad

Tu equipo ha creado una librería de *Deep Learning* propietaria que no es ni TensorFlow ni PyTorch. ¿Cómo resuelves el desafío de la **Reproducibilidad** y la **Escalabilidad** para entrenar este modelo en la nube sin tener que reescribir todo el código?

Escenario B: MLOps y Desafíos Fundamentales

Pregunta sobre Data-Serving Skew

Tu modelo de predicción de *churn* (abandono de clientes) fue muy preciso durante las pruebas, pero empezó a fallar tan pronto como fue desplegado. El análisis indica que la *feature* "antigüedad del cliente" se calculó de forma diferente en el entorno de entrenamiento y en el entorno de producción. ¿Qué desafío de MLOps es este y qué servicio de SageMaker se utiliza para prevenirlo?

Pregunta sobre Plataformas Gestionadas

¿Cuál es la principal diferencia entre usar una **Instancia EC2** (IaaS) con una GPU para el entrenamiento manual y usar una plataforma **SageMaker Training Job** (PaaS) para el mismo entrenamiento, en términos de esfuerzo de Ingeniería?

Preguntas Clave de la Sesión 9 (Despliegue, Monitorización y Orquestación)

Bloque 1: Despliegue e Inferencia (SageMaker Endpoints)

Pregunta de Componentes y Reproducibilidad

El artefacto **model.tar.gz** es el resultado de un *Training Job* exitoso y la unidad para el despliegue. ¿Cuáles son los tres archivos esenciales que contiene y por qué el **requirements.txt** es crucial para la **Reproducibilidad** en el *endpoint*?

Pregunta de Caso de Uso de Tareas Pesadas

Una firma de arquitectura necesita procesar **500 modelos 3D de alta resolución** cada noche. Cada uno tarda 15 minutos en analizarse. ¿Qué modo de inferencia es el más adecuado y por qué **no** usaría el *Real-time Endpoint*?

Pregunta de Variantes de Modelos

¿Qué característica de los **SageMaker Real-time Endpoints** permite a las organizaciones realizar un **A/B Testing** de un nuevo modelo (Candidato B) contra la versión actual (Modelo A) sin cambiar la dirección de la API?

Bloque 2: Monitorización, Orquestación y Comparativa Cloud

Pregunta de Model Monitor y Data Drift

¿Cómo utiliza **SageMaker Model Monitor** la **Línea Base (*Baseline*)** para detectar el **Data Drift** en producción?

Pregunta de Orquestación y Gobernanza

¿Cuál es el rol del `ConditionStep` y el `RegisterModelStep` dentro de un **SageMaker Pipeline**, y qué desafío de MLOps resuelven juntos?

Pregunta de Estrategia de Plataforma Cloud

Un equipo valora una interfaz de usuario **muy unificada** y necesita la infraestructura de **TPUs** para *deep learning* a gran escala. ¿Qué plataforma, alternativa a SageMaker, es la más adecuada?

Actividad de Gamificación: Sesión 9 (Despliegue, Monitorización y Orquestación)

Escenario A: Despliegue de Alto Tráfico y Baja Latencia
Descripción: Eres una gran plataforma de *e-commerce* y necesitas un modelo de recomendación que responda en **milisegundos** a **cientos de miles de peticiones constantes** por hora, 24/7. Tu tráfico es alto y constante.

- **Opciones:** 1. Inferencia Asíncrona, 2. Real-time Endpoint (Aprovis.), 3. Batch Transform.

Escenario B: MLOps y Despliegue Condicional Automatizado
Descripción: Has entrenado un nuevo modelo, y para el MLOps automatizado, solo debe pasar a despliegue si su **F1-score es superior al 85%**. Si no lo es, el *pipeline* debe detenerse. ¿Qué *Step* de **SageMaker Pipelines** se encarga de verificar esta métrica y controlar el flujo del DAG?

- **Opciones:** 1. EvaluationStep, 2. ConditionStep, 3. Model Monitor.

Escenario C: Eficiencia de Costes e Intermitencia
Descripción: Estás desarrollando un modelo interno para RRHH que predice la rotación de personal, pero solo se ejecuta de forma **esporádica o intermitente** (una o dos veces al mes). Necesitas pagar solo por el uso activo.

- **Opciones:** 1. Real-time Endpoint (Aprovis.), 2. Serverless Inference, 3. Batch Transform.

Escenario D: Detección de Drift en Producción

Descripción: Tu modelo de clasificación de spam ha sido desplegado, pero las nuevas campañas de *phishing* han cambiado la distribución de los correos de entrada (**Data Drift**). ¿Qué servicio **compara automáticamente** el tráfico en vivo capturado con la **línea base** de entrenamiento para generar una alerta en CloudWatch?

- **Opciones:** 1. SageMaker Debugger, 2. SageMaker Model Monitor, 3. CloudWatch Alarms.

Escenario E: Procesamiento de Tareas Pesadas (Flujo Asíncrono)

Descripción: Una aplicación necesita analizar documentos extensos de hasta **1 GB de tamaño** y acepta que la predicción tarde unos minutos, siempre y cuando reciba una **notificación** cuando el resultado esté depositado en S3.

- **Opciones:** 1. Real-time Endpoint, 2. Batch Transform, 3. Inferencia Asíncrona.