Manual de Estudio: AWS Certified AI Practitioner

Dominio 2.3: Infraestructura y Tecnologías de AWS para IA Generativa

Material de Preparación Detallado

1. Ventajas de la Infraestructura de AWS

Crear aplicaciones de IA generativa en AWS ofrece beneficios estratégicos clave frente a soluciones locales o tradicionales:

- **Accesibilidad y Eficiencia:** Reducción de las barreras de entrada para empresas sin grandes laboratorios de investigación.
- **Rapidez de Comercialización:** Uso de modelos pre-entrenados para evitar el ciclo largo de entrenamiento desde cero.
- **Aprendizaje por Transferencia (*Transfer Learning*):** Proceso de refinar un modelo pre-entrenado con un conjunto de datos pequeño. El algoritmo ya "sabeïnformación general y solo necesita aprender a mapear elementos a los nuevos datos, ahorrando tiempo y costes significativos.

2. La Pila de IA Generativa de AWS (3 Capas)

AWS divide su oferta tecnológica en tres niveles de abstracción según la necesidad del cliente:

1. **Capa Inferior (Infraestructura):** Herramientas para crear y entrenar LLMs. Incluye hardware especializado con la mejor relación precio-rendimiento:
   - **AWS Trainium:** Acelerador para el entrenamiento de modelos.
   - **AWS Inferentia:** Acelerador optimizado para la inferencia (uso del modelo).
   - **Instancias con GPU:** Familias P4, P5, G5 y G6 de Amazon EC2.
2. **Capa Intermedia (Herramientas):** Acceso a modelos y plataformas para construir y escalar sin gestionar la infraestructura base. El servicio estrella es **Amazon Bedrock**.
3. **Capa Superior (Aplicaciones):** Aplicaciones terminadas que consumen FMs para tareas como escribir código (Amazon Q Developer) o arquitecturas tipo **RAG** (Generación Aumentada por Recuperación).

3. Seguridad y Privacidad en IA Generativa

La seguridad es la prioridad número uno en AWS. Se enfoca en proteger la tríada: **Entrada, Modelo y Salida**.

3.1. Protección de Infraestructura

- **AWS Nitro System:** Hardware y firmware especializado que aplica restricciones de seguridad para garantizar que nadie (ni siquiera el personal de AWS) pueda acceder a los datos de las cargas de trabajo en instancias EC2.
- **Cifrado y MFA:** Implementación obligatoria de cifrado de datos y autenticación multifactor.

3.2. Vulnerabilidades Específicas de la IA

- **Inyección de Peticiones (*Prompt Injection*):** Intentos de engañar al modelo mediante entradas maliciosas para que ignore sus restricciones de seguridad.
- **Envenenamiento de Datos (*Data Poisoning*):** Manipular los datos de entrenamiento para que el modelo aprenda patrones incorrectos o maliciosos.
- **Inversión de Modelos:** Intentos de extraer datos confidenciales del dataset de entrenamiento a partir de las respuestas del modelo.

4. Servicios Clave de AWS para GenAI

4.1. Amazon Bedrock (Servicio Administrado)

Permite acceder a Modelos Fundacionales (FM) de AWS (Amazon Titan) y de terceros (Anthropic, Meta, Cohere, Stability AI) mediante una API.

- **Zonas de Juegos (*Playgrounds*):** Entornos para experimentar con diferentes modelos y ajustar parámetros de inferencia antes de programar.
- **PartyRock:** Una herramienta basada en Bedrock para aprender técnicas de *prompting* creando aplicaciones sencillas (listas de reproducción, recetas, etc.).
- **Importación de Ponderaciones:** Capacidad de traer pesos personalizados de modelos entrenados fuera de Bedrock.

4.2. Amazon SageMaker JumpStart

Un centro de modelos dentro de SageMaker que ayuda a encontrar proyectos preconstruidos, algoritmos y soluciones basadas en mejores prácticas. Requiere la gestión de GPUs para el refinamiento y despliegue, por lo que es vital monitorizar costes y eliminar puntos de enlace (*endpoints*) cuando no se usen.

5. Modelos de Precios y Costes

Existen dos formas principales de pagar por servicios de GenAI:

1. **Alojamiento de Infraestructura:** Pagas por los recursos de computación (EC2, instancias de SageMaker) que ejecutan el modelo. Eres responsable de la gestión de la capacidad.

2. **Precio por Tokens:** Pagas por la cantidad de unidades de información (texto, píxeles) procesadas. Es el modelo típico de las llamadas a API en **Amazon Bedrock** (Pago por uso).

6. Glosario Estratégico

- **CAF-AI (*AWS Cloud Adoption Framework for AI*):** Marco de trabajo guía para debatir estrategias de IA con socios y compañeros.
- **Bases de Datos Vectoriales:** Almacenan datos como **incrustaciones (embeddings)** para realizar búsquedas semánticas avanzadas de forma eficiente.
- **Alta Disponibilidad:** La infraestructura global de AWS (Regiones y Zonas de Disponibilidad) garantiza que las aplicaciones de IA sean tolerantes a fallos.