Guía de Arquitectura Avanzada y Escenarios

AWS Certified AI Practitioner (AIF-C01)

Material de Refuerzo de Alta Intensidad

1. Comparativa Técnica Profunda: Bedrock vs. SageMaker

En el examen, te presentarán empresas con diferentes niveles de madurez técnica. Debes saber mapear el servicio exacto según la necesidad.

Criterio	Amazon Bedrock	Amazon SageMaker
Infraestructura	**Serverless por defecto**. No gestionas parches ni escalado de instancias.	**Basado en instancias**. Requiere elegir tipos de instancia (ML.m5, ML.g5, etc.).
Modelos	Acceso a FMs líderes (Claude, Llama, Mistral, Titan) vía **API unificada**.	Permite usar **cualquier modelo** de Hugging Face o algoritmos propios en contenedores Docker.
Personalización	Soporta *Fine-tuning* y *Continued Pre-training* de forma gestionada.	Control total sobre el bucle de entrenamiento, hiperparámetros y optimizadores.
Seguridad	Aislamiento lógico. Los datos no salen de la región ni se usan para entrenar modelos base.	Aislamiento de red profundo mediante VPC, grupos de seguridad y PrivateLink dedicado.
Ideal para...	Aplicaciones de GenAI que requieren **Time-to-Market rápido**.	Proyectos de **Investigación y Desarrollo**, ML tradicional y personalización extrema.

2. Mecánica de Precios en Amazon Bedrock

AWS no solo te preguntará cuánto cuesta, sino **cómo se factura** según la carga de trabajo.

2.1. 1. On-Demand (Bajo Demanda)

- **Cálculo:** Se factura por cada 1,000 tokens procesados.

- **Diferenciación:** Los tokens de **Entrada** suelen ser más baratos que los de **Salida** (generación).
- **Escenario:** Desarrollo inicial, pruebas y aplicaciones con picos de tráfico impredecibles.

2.2. 2. Provisioned Throughput (Capacidad Aprovisionada)

- **Cálculo:** Compras Ünidades de Modelo"(Model Units) con un compromiso temporal (1 o 6 meses).
- **Garantía:** Asegura una **tasa de transferencia constante** (tokens por minuto) sin variaciones de latencia.
- **Escenario:** Aplicaciones críticas en producción con tráfico constante que no pueden permitirse retardos por çola".

2.3. 3. Model Customization (Costo de Refinamiento)

- **Cálculo:** Se cobra por el almacenamiento del modelo personalizado y por el proceso de entrenamiento (*hours/tokens*).

3. Ecosistema Amazon Q: El Asistente Inteligente

Es común que el examen mezcle .ªmazon Q Businessçon .ªmazon Q Developer". Aquí está la distinción definitiva:

Amazon Q Developer

Para quién: Desarrolladores e Ingenieros de AWS.
Dónde actúa: En el IDE y la Consola de AWS.
Función principal: Escribir código, convertir lenguajes (ej. Java 8 a Java 17), depurar errores de red y sugerir arquitecturas en la nube.

Amazon Q Business

Para quién: Empleados de cualquier departamento (Ventas, HR, Legal).
Dónde actúa: En portales web corporativos.
Función principal: Responder preguntas basadas en los datos internos de la empresa (SharePoint, Slack, Salesforce) respetando los permisos de acceso originales.

Amazon Q in QuickSight

Para quién: Analistas de datos y Directivos.
Función principal: Crear dashboards visuales a partir de preguntas en lenguaje natural ("Muéstrame el ROI por país").

4. Pipeline Técnico de RAG (Bases de Conocimiento)

Si la pregunta pide los pasos lógicos para configurar **Amazon Bedrock Knowledge Bases**, el orden es:

1. **Data Source (S3):** Subir los documentos (*raw data*) a un bucket.
2. **Chunking strategy:** Dividir el texto (ej. trozos de 300 tokens con 10 % de solapamiento).
3. **Embedding Model:** Seleccionar un modelo (ej. *Titan Text Embeddings v2*) para transformar texto en vectores.
4. **Vector Store Setup:** Elegir dónde indexar los vectores (**OpenSearch Serverless** es la recomendación *managed*).
5. **Ingestion Job:** Sincronizar el origen de datos para procesar y almacenar los vectores.

5. Servicios "No-Code"para Usuarios de Negocio

Si el escenario describe a un usuario que **no sabe programar** pero quiere hacer ML:

- **Amazon SageMaker Canvas:** Interfaz visual de arrastrar y soltar para predecir valores de negocio (ej. rotación de clientes).
- **Amazon Bedrock Playgrounds:** Interfaz web para probar modelos generativos y ajustar temperatura/parámetros sin escribir una sola línea de código.

6. Seguridad de Grado Empresarial en GenAI

- **AWS PrivateLink:** Permite que el tráfico entre tu VPC y Amazon Bedrock sea **totalmente privado** (no toca el internet público).
- **AWS KMS:** Tú gestionas las claves para cifrar los modelos personalizados y las bases de conocimiento.
- **Amazon Macie:** Detecta si has subido accidentalmente documentos con números de tarjeta de crédito (PII) a tu base de conocimientos de RAG.