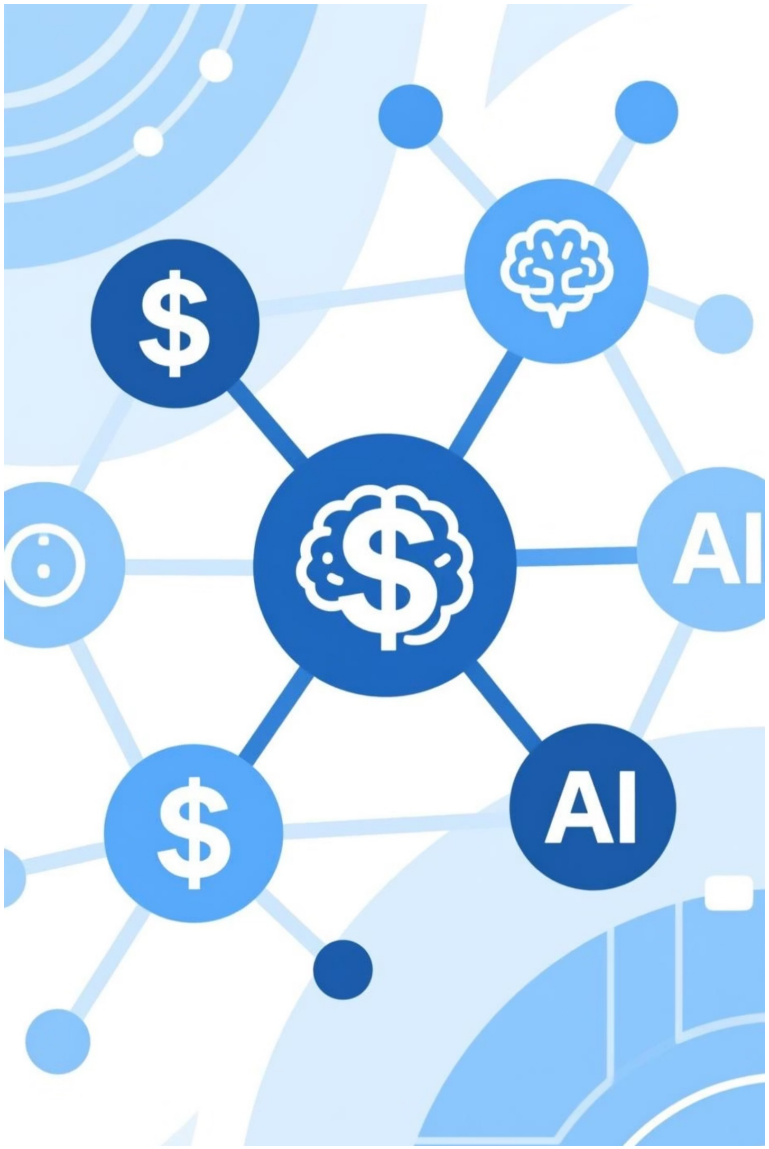


Cloud Computing para Inteligencia Artificial

Sesión 12: Optimización de Costes y Gestión Financiera (FinOps) en Cloud para IA

Prof. [Nombre del Profesor] • Universidad [Nombre] • [Fecha]

Objetivos de Aprendizaje

Al finalizar esta sesión, serás capaz de:

Comprender

Los principios fundamentales de FinOps y su importancia en proyectos de IA

Diferenciar

Los modelos de precios de los principales proveedores de nube para cargas de trabajo de IA

Identificar

Los principales generadores de costes en una infraestructura de IA en la nube

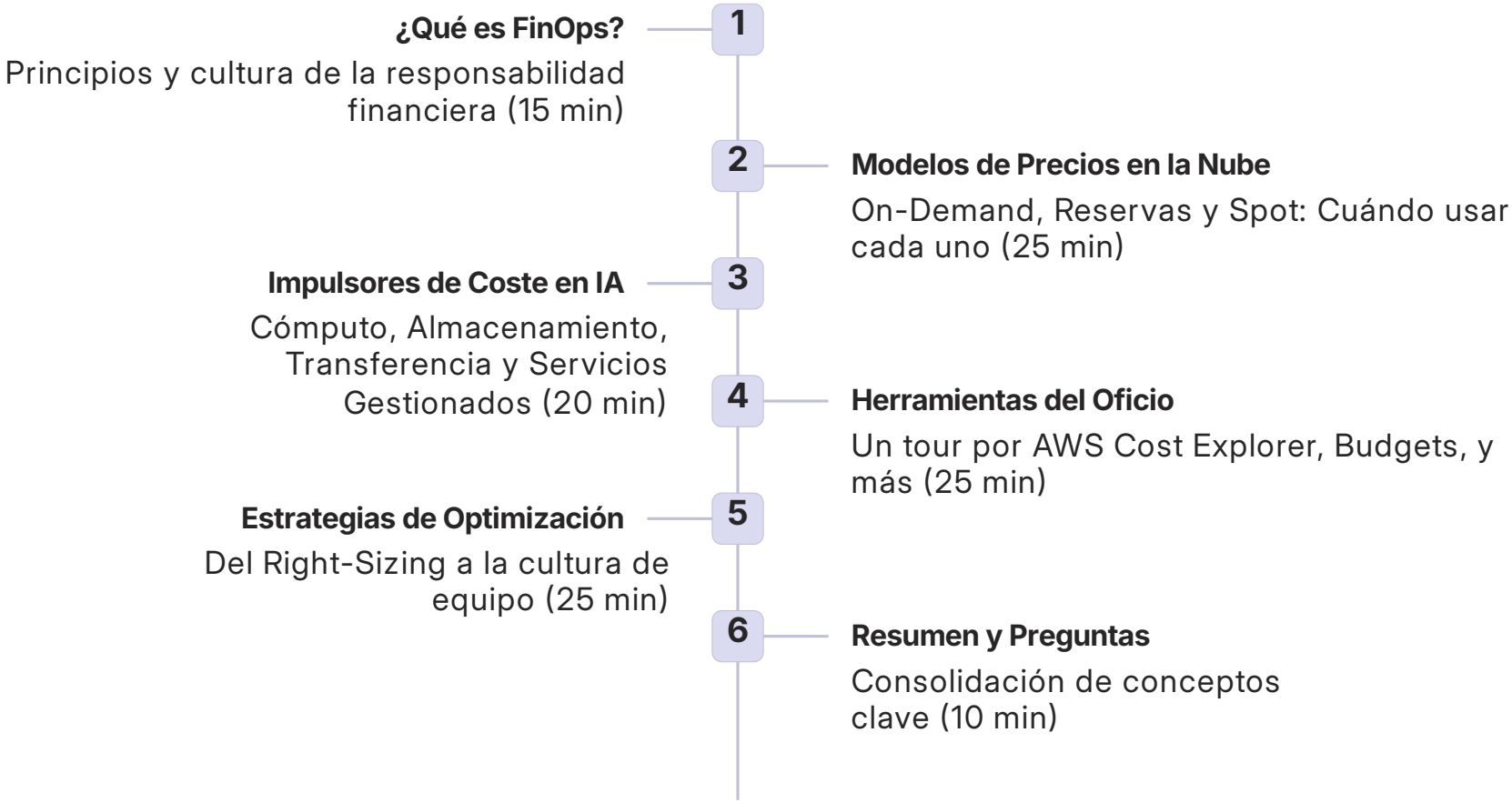
Utilizar

Las herramientas nativas de la nube para monitorizar, analizar y gestionar los costes

Aplicar

Estrategias prácticas para optimizar los costes de entrenamiento e inferencia de modelos de ML

Agenda de la Sesión



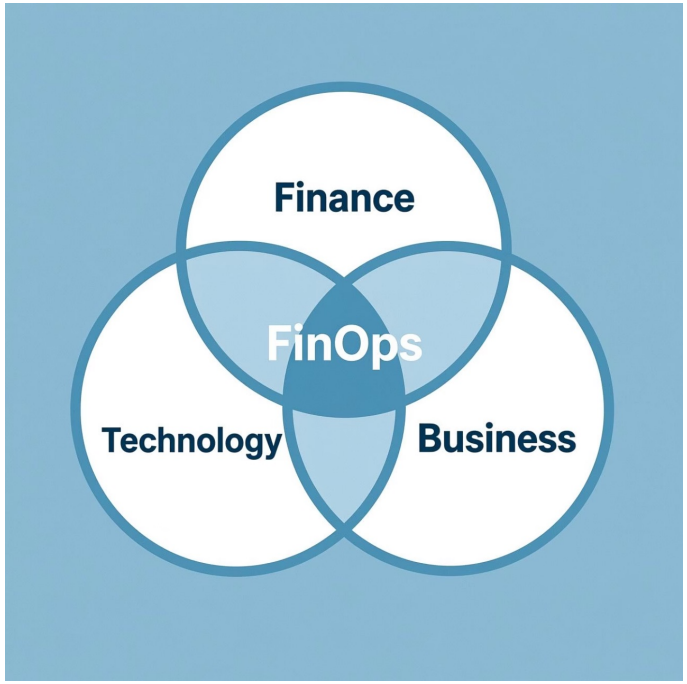
FinOps - Más Allá de "Apagar lo que no usas"

FinOps (Cloud Financial Operations) es una disciplina y práctica cultural que impulsa la responsabilidad financiera y maximiza el valor de negocio del gasto en la nube.

No se trata solo de ahorrar dinero, sino de **ganar más dinero** tomando decisiones de ingeniería informadas por su impacto financiero.

> **La Ecuación Clave:** Valor de Negocio = (Rendimiento del Modelo * Velocidad de Iteración) / Coste de Infraestructura

Piensa en FinOps como llevar la agilidad de DevOps al ámbito financiero de la nube. En lugar de un presupuesto fijo anual, gestionamos un gasto variable y dinámico.



Los Principios Clave de FinOps

1

Colaboración Interfuncional

Los equipos de Finanzas, Ingeniería (IA/ML) y Negocio deben hablar el mismo idioma y compartir objetivos. Los ingenieros deben entender el impacto de su código en la factura.

2

Propiedad y Responsabilidad

Los equipos que construyen y ejecutan servicios son responsables de su propio gasto en la nube. *"Tú lo construyes, tú lo ejecutas, tú lo costeas"*.

3

Decisiones Basadas en el Valor

Cada euro gastado en la nube debe justificarse con un retorno o valor para el negocio. ¿Entrenar este modelo una hora más mejorará la precisión lo suficiente como para justificar el coste de la GPU?

4

Visibilidad y Transparencia

Los equipos necesitan acceso fácil y comprensible a sus datos de costes casi en tiempo real para tomar decisiones informadas.

5

Optimización Continua

La gestión de costes no es un proyecto puntual, es un proceso iterativo y constante.

El Menú de Precios de la Nube

La flexibilidad de la nube se refleja en sus modelos de precios. Elegir el incorrecto es como pagar un menú de degustación cuando solo querías un café.

Pago por Uso (On-Demand)

Máxima flexibilidad, sin compromisos. Ideal para empezar, para cargas de trabajo impredecibles o de corta duración.

Compromisos de Uso (Reservas)

Descuentos significativos (hasta 72%) a cambio de un compromiso de uso (1 o 3 años). Ideal para cargas de trabajo estables y predecibles.

Capacidad Excedente (Spot)

Descuentos masivos (hasta 90%), pero la instancia puede ser interrumpida con poco aviso. Ideal para cargas de trabajo tolerantes a fallos.

Pago por Uso (On-Demand) - La Opción Flexible

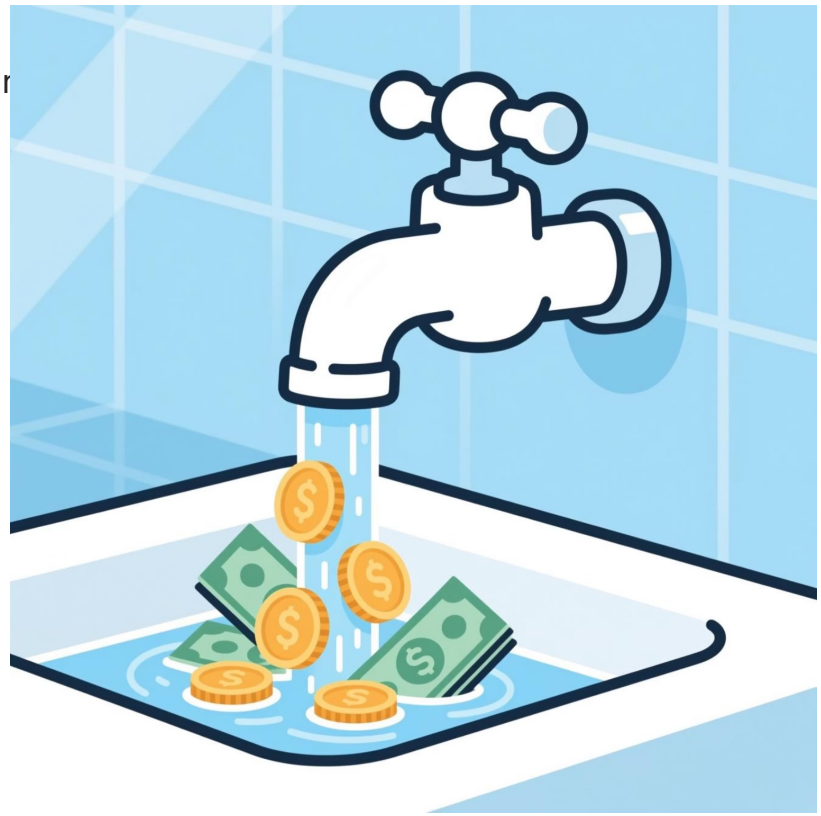
Cómo funciona:

Pagas por la capacidad de cómputo por segundo o por hora, sin

Ideal para:

- Desarrollo, pruebas y experimentos de modelos de IA donde no conoces la duración
- Aplicaciones con picos de demanda a corto plazo e impredecibles
- La primera vez que despliegas un endpoint de inferencia para medir su demanda real

⊗ **La Trampa:** Es el modelo más caro por hora. Dejar recursos On-Demand encendidos por inercia es la principal fuente de gasto inesperado.



Ejemplo: Lanzas una instancia p4d.24xlarge en AWS para un prototipado rápido. La usas 45 minutos y la apagas. Solo pagas por esos 45 minutos.

Instancias Reservadas y Savings Plans (AWS)

Te comprometes a usar una cierta cantidad de cómputo durante 1 o 3 años a cambio de un gran descuento.

Standard RIs

El mayor descuento, pero ligado a una familia de instancias específica (ej: m5) en una región. Menos flexible.

Convertible RIs

Menor descuento, pero permite cambiar la familia de instancias, el sisten

EC2 Instance Savings Plans

Descuento a cambio de un compromiso de gasto por hora (ej: $10/hora) en una familia de instancias en una región. Más flexible que las RIs.

Compute Savings Plans

El más flexible. El descuento se aplica a cualquier uso de cómputo (EC2, Fargate, Lambda) en cualquier región.

Comparativa de Compromisos: Azure y GCP

Azure

Reservations

Similar a las RIs de AWS. Compras capacidad específica para un periodo de 1 o 3 años.

Azure Savings Plans for compute

Similar a los Compute Savings Plans de AWS. Te comprometes a un gasto por hora y obtienes descuentos en una amplia gama de servicios de cómputo.

Google Cloud Platform (GCP)

Committed Use Discounts (CUDs)

- **Basados en recursos:** Te comprometes a una cantidad de vCPUs y memoria en una región. Similar a las RIs.
- **Basados en gasto:** Te comprometes a un gasto por hora. Muy flexible.

Sustained Use Discounts (SUDs)

¡Automáticos! GCP aplica descuentos automáticamente cuanto más tiempo se ejecute una VM en un mes. No requiere compromiso. Es un beneficio único de GCP.

Instancias Spot - El Secreto Mejor Guardado para IA

¿Qué son?

Son capacidad de cómputo no utilizada que los proveedores de nube venden a un precio muy bajo.

El Trato:
Obtienes descuentos de hasta el **90%** sobre el precio On-Demand.

La "Letra Pequeña":

El proveedor puede reclamar esa capacidad en cualquier momento con un aviso de 2 minutos (AWS) o 30 segundos (GCP). Tu instancia es interrumpida.

¿Por qué son perfectas para IA?

- **Entrenamiento de modelos:** Muchos frameworks (TensorFlow, PyTorch) soportan checkpoints. Si la instancia se interrumpe, puedes reanudar el entrenamiento desde el último checkpoint en una nueva instancia Spot.
- **Procesamiento por lotes (Batch processing):** Trabajos de ETL, preprocesamiento de datos.
- Simulaciones y renderizado.

Estrategias para Usar Spot con Éxito

No te cases con una instancia

Solicita múltiples tipos de instancias (ej: p3.2xlarge, p3.8xlarge, g4dn.xlarge). Si un tipo no está disponible, tu clúster puede obtener otro.

Diversifica entre Zonas de Disponibilidad (AZs)

La disponibilidad de Spot varía por AZ. Lanza instancias en múltiples AZs para aumentar la probabilidad de obtener capacidad.

Utiliza servicios gestionados que lo simplifican

- **AWS SageMaker Managed Spot Training:** Con un simple parámetro (use_spot_instances=True), SageMaker gestiona los checkpoints y la reanudación del entrenamiento por ti. Puede ahorrar hasta un 90% en costes de entrenamiento.
- **AWS EC2 Fleet / Auto Scaling Groups:** Permite definir una composición de flota con un porcentaje de instancias On-Demand (para la base) y un porcentaje de Spot (para escalar).

¿De Dónde Vienen los Cargos? Principales Impulsores de Coste

Una solución de IA no es un solo servicio, es un ecosistema de recursos. Entender cada componente es clave para controlar el coste total.

Cómputo (compute)

El "motor". Las instancias (CPU/GPU) donde se entrenan los modelos y se ejecutan las inferencias. **Suele ser >70% del coste total.**

Almacenamiento (storage)

El "combustible". Donde viven tus datasets, modelos entrenados, logs y backups.

Transferencia de Datos (data transfer)

Las "tuberías". Mover datos dentro y fuera de la nube.

Servicios Gestionados de IA/ML

La "inteligencia". Plataformas como SageMaker, Azure ML, o APIs cognitivas.

El Coste del Cómputo: El Gigante Silencioso

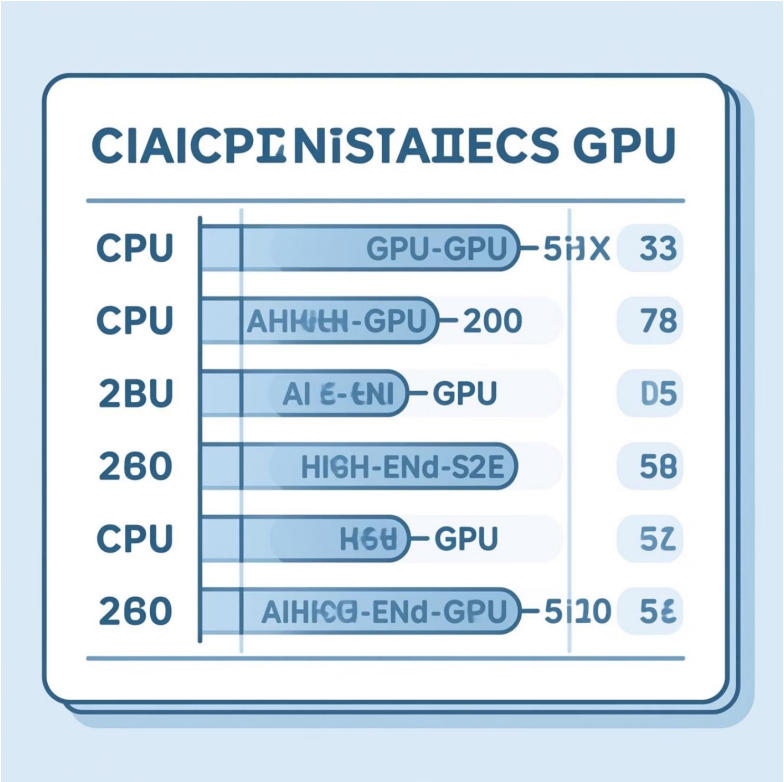
⚠ Una p4d.24xlarge de AWS puede costar >$30/hora On-Demand. ¡Eso es ~$260,000 al año si se ejecuta 24/7!

Factores que influyen:

- **Tipo de instancia:** GPU vs. CPU, generación, tamaño.
- **Duración del trabajo:** Un entrenamiento que dura 100 horas vs. 10 horas.
- **Escalado:** El número de instancias para entrenamiento distribuido o el número de réplicas en un endpoint de inferencia.

Pregunta Clave:

¿Realmente necesito una A100 para este experimento, o una T4 sería suficiente?



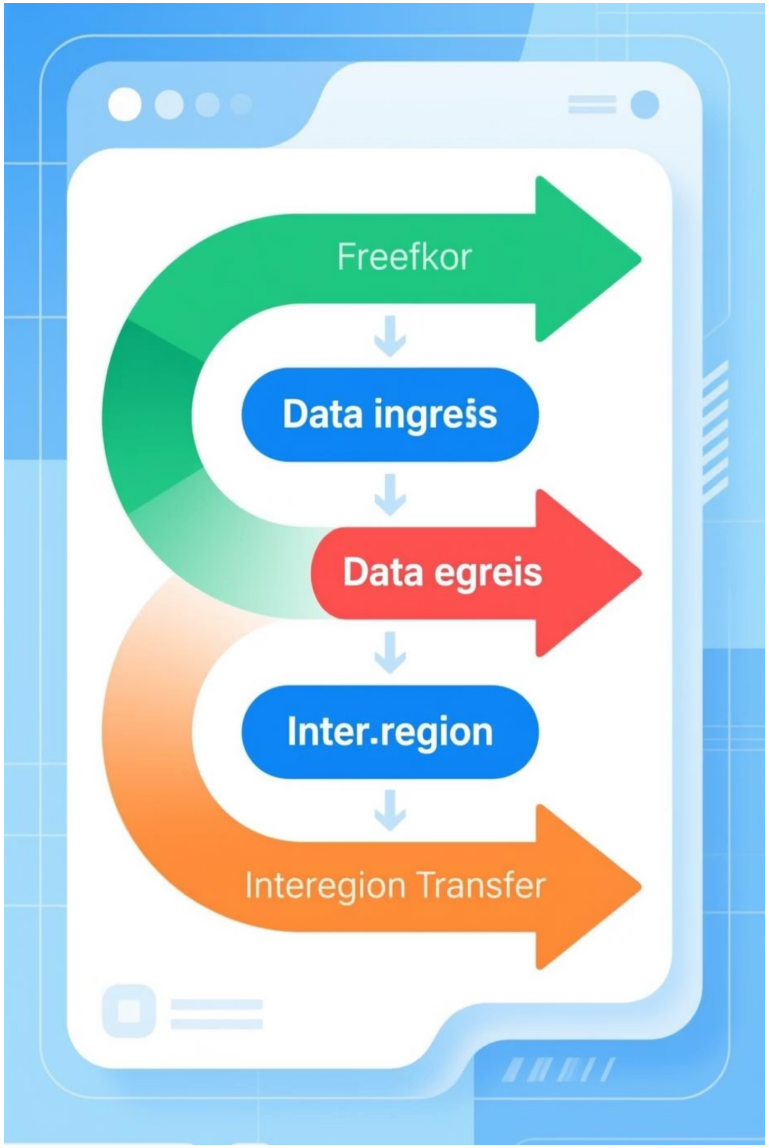
Almacenamiento y Transferencia: Las Fugas Lentas

Almacenamiento

- **Volumen de datos:** Datasets de Terabytes, modelos de Gigabytes, logs acumulados.
- **Clase de almacenamiento:** AWS S3 Standard (acceso frecuente) es ~10x más caro que S3 Glacier Deep Archive (archivado a largo plazo). Elegir la clase correcta es crucial.
- **Snapshots y Backups:** Son importantes, pero los snapshots olvidados pueden acumular costes significativos.

Transferencia de Datos

- **Salida de datos (Egress):** ¡CUIDADO! Mover datos **FUERA** de la nube a internet suele tener un coste elevado. Los datos que entran (Ingress) son generalmente gratuitos.
- **Transferencia entre regiones:** También tiene coste.
- **NAT Gateways:** Un servicio muy útil pero que puede generar costes altos si no se monitoriza el tráfico que pasa a través de él.



Servicios Gestionados de IA y Monitorización

AWS SageMaker (Ejemplo)

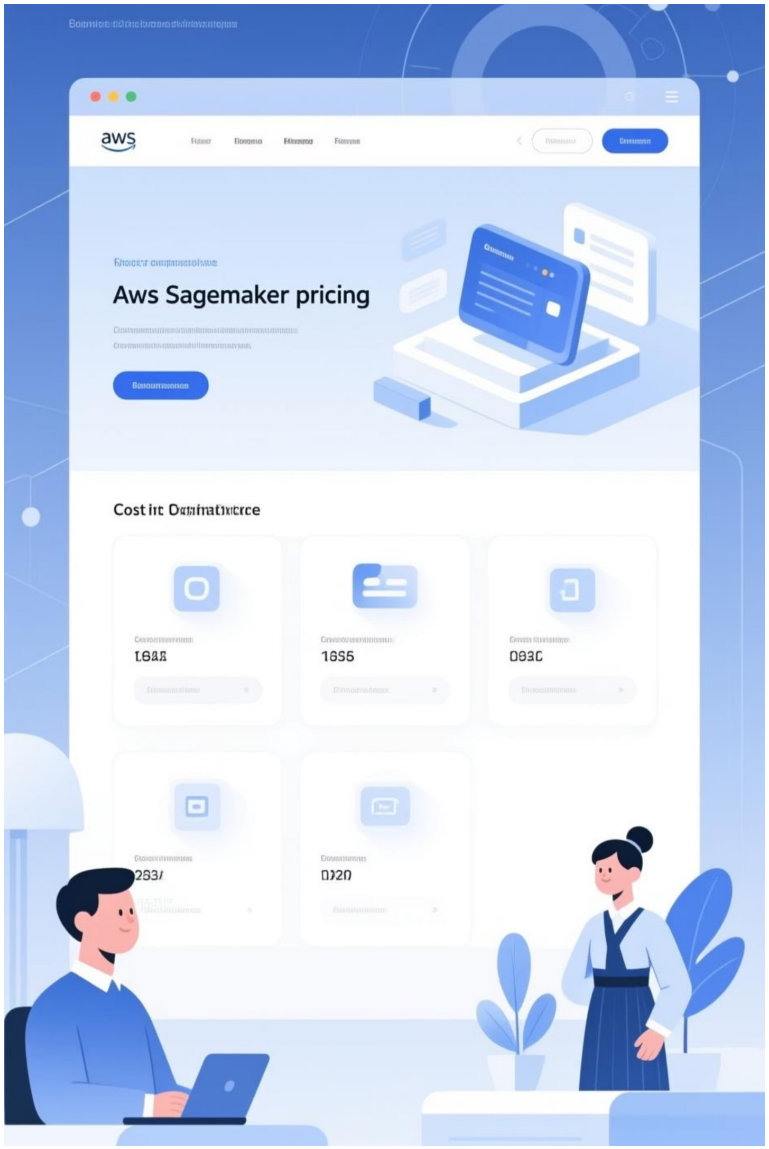
- **Notebooks (Studio/Instancias):** Pagas por el tipo y duración de la instancia que ejecuta tu notebook. ¡Apágala cuando no la uses!
- **Training Jobs:** Pagas por el tipo y duración de las instancias de entrenamiento.
- **Inference Endpoints:** Pagas por el tipo y duración de las instancias que sirven tu modelo, **independientemente de si reciben tráfico o no.**

APIs Cognitivas

El coste se basa en el número de llamadas o la cantidad de datos procesados (imágenes, texto, video).

Logging y Monitorización

El almacenamiento de logs y la creación de métricas personalizadas pueden sumar costes si no se gestionan (ej. logs con nivel de DEBUG en producción).



GAMIFICACIÓN - Elige la Arquitectura Correcta

> "Estás liderando un proyecto para entrenar un modelo de Procesamiento de Lenguaje Natural (BERT) desde cero. El dataset tiene 500GB. El entrenamiento se estima que durará 8 días en un clúster de 8 instancias p4d.24xlarge. El proceso puede reanudarse desde checkpoints guardados cada 6 horas. El presupuesto es muy ajustado."

¿Qué combinación de modelos de precios elegirías y por qué?

Opción A

100% On-Demand. (Máxima seguridad, coste altísimo).

Opción B

100% Spot. (Ahorro máximo, riesgo muy alto de interrupciones y retrasos).

Opción C

100% Savings Plan a 1 año. (Gran descuento, pero ¿qué pasa después de los 8 días de entrenamiento? Compromiso a largo plazo para una necesidad a corto plazo).

Opción D

Una flota mixta: 1 instancia On-Demand/Reservada (para el nodo principal o "worker 0") y 7 instancias Spot para el resto del clúster. (Equilibrio entre coste y fiabilidad).

Visibilidad es Control: Herramientas de Gestión de Costes

No puedes optimizar lo que no puedes ver. El primer paso es entender tu factura.

Nivel 1: Visualización y Alerta

AWS Cost Explorer, AWS Budgets. Para una visión general y alertas tempranas.

Nivel 2: Análisis Profundo

AWS Cost and Usage Report (CUR). Para análisis forenses y detallados.

Nivel 3: Recomendaciones Proactivas

AWS Trusted Advisor, AWS Cost Anomaly Detection. Para optimización guiada por la nube.

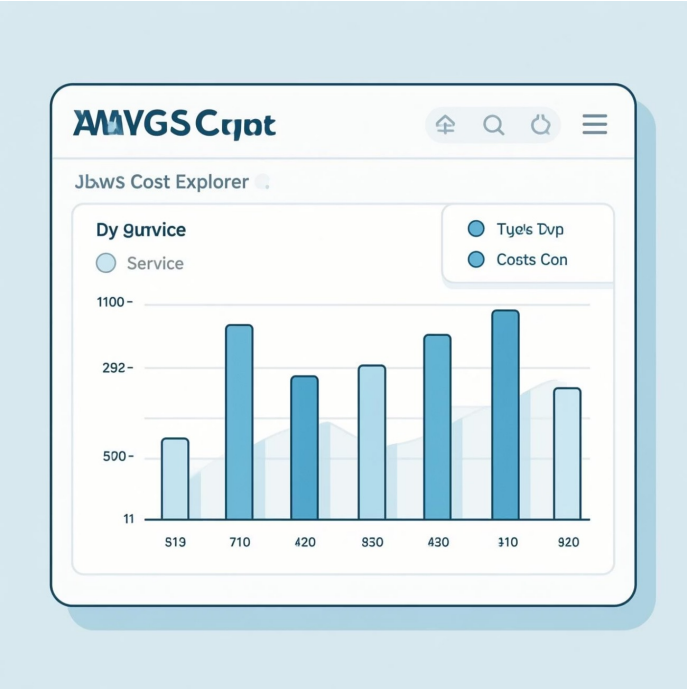
AWS Cost Explorer - Tu GPS Financiero

¿Qué es?

Una interfaz visual para explorar tus costes y uso.

Funcionalidades Clave:

- **Visualización Histórica:** Observa tus costes de los últimos 12 meses.
- **Filtrado y Agrupación:** Analiza costes por servicio (EC2, S3, SageMaker), por región, por cuenta, ¡y por etiquetas!
- **Previsiones (Forecasting):** Basado en tu historial, Cost Explorer puede predecir tu factura a final de mes.
- **Informes Guardados:** Crea y guarda informes personalizados para tus proyectos de IA.



Caso de Uso para IA:

Agrupar costes por la etiqueta `project:bert-training` para ver exactamente cuánto costó ese experimento.

AWS Budgets - Tu Perro Guardián

¿Qué es?

Te permite establecer umbrales de coste y recibir alertas cuando se superan.

Tipos de Presupuestos:

- **Coste:** Alerta cuando el gasto total (o filtrado) supera los $X.
- **Uso:** Alerta cuando usas más de X horas de un tipo de instancia.

¡Superpoder! Budgets Actions

No solo te alerta, sino que puede **actuar**.

Ejemplo:

Si un presupuesto para un entorno de desarrollo supera el 100%, una Budget Action puede aplicar automáticamente una política IAM que deniegue el permiso para lanzar nuevas instancias de GPU, evitando una catástrofe financiera.

CUR y Detección de Anomalías - Nivel Experto

Cost and Usage Report (CUR)

- El informe más detallado posible. Es un fichero CSV o Parquet entregado a un bucket S3.
- Contiene cada cargo individual con una granularidad horaria.
- **¿Para qué?** Para análisis muy avanzados usando herramientas como Amazon Athena (SQL sobre S3) y QuickSight (BI) para crear dashboards personalizados.

AWS Cost Anomaly Detection

- Utiliza Machine Learning para aprender tu patrón de gasto normal.
- Te alerta automáticamente si detecta un gasto inusual que no se ajusta a tu patrón, incluso si no ha superado el umbral de tu presupuesto.
- **Ejemplo:** Detecta que un job de ETL empezó a escribir logs excesivos a S3, generando un coste inesperado.

El Poder del Etiquetado (Tagging)

¿Qué es?

Las etiquetas son metadatos (pares clave-valor) que asignas a tus recursos de AWS (instancias EC2, buckets S3, endpoints de SageMaker).

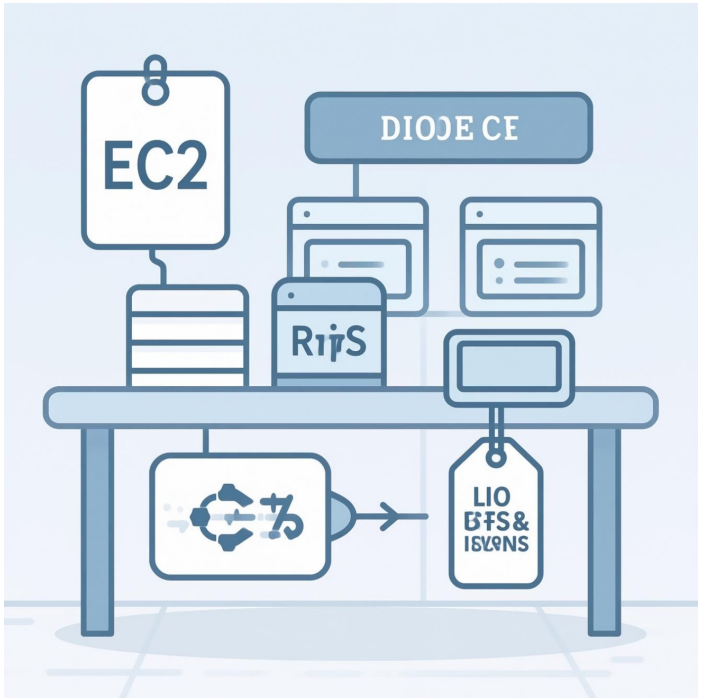
¿Por qué es CRÍTICO?

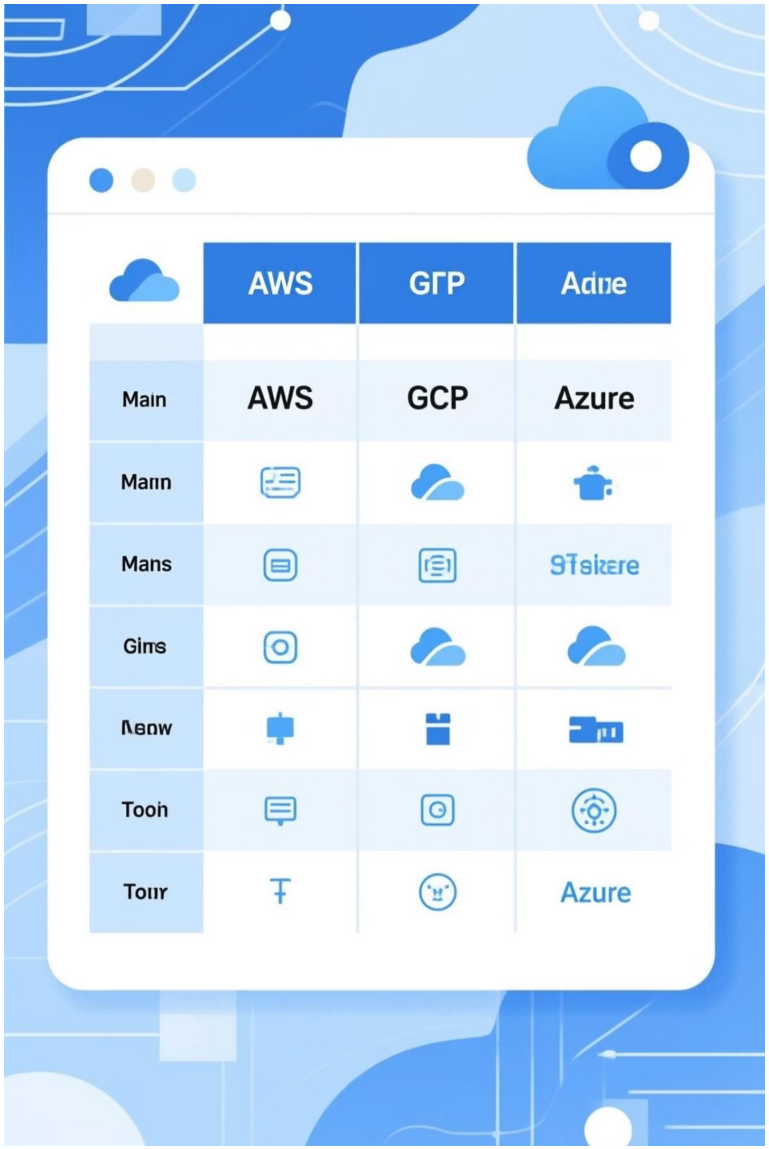
Sin etiquetas, tu factura es un bloque monolítico. Con etiquetas, puedes desglosar los costes por:

- `project: face-recognition-v2`
- `environment: development, staging, production`
- `team: data-science-alpha`
- `owner: profesor-x`

Estrategia de Tagging:

La clave es la **consistencia**. Define una política de etiquetado obligatoria para toda la organización y usa herramientas como AWS Service Catalog o AWS Organizations para aplicarla.





Comparativa de Herramientas: GCP y Azure

Los conceptos son universales, solo cambian los nombres.

Google Cloud Platform (GCP)

- **Herramienta Principal:** Cloud Billing Reports (similar a Cost Explorer).
- **Análisis Profundo:** Export Billing Data to BigQuery (equivalente a CUR + Athena, pero más integrado).
- **Etiquetado:** Se llaman "Labels".
- **Recomendaciones:** Integradas directamente en la consola (Recommender API).

Microsoft Azure

- **Herramienta Principal:** Azure Cost Management and Billing. Ofrece análisis, presupuestos y alertas.
- **Análisis Profundo:** Export Cost Data.
- **Etiquetado:** Se llaman "Tags".
- **Recomendaciones:** Azure Advisor.

Estrategia 1: Right-Sizing - No Mates Moscas a Cañonazos

Definición:

Right-sizing es el proceso de analizar el rendimiento y la capacidad de tus recursos y ajustarlos para que coincidan con la carga de trabajo real, evitando el sobreaprovisionamiento.

En IA, esto significa:

- **CPU/Memoria:** ¿Estás usando una instancia m5.8xlarge para un notebook de Jupyter donde solo usas el 10% de la CPU? Monitoriza con CloudWatch Metrics.
- **GPU:** ¿Tu código de entrenamiento utiliza eficientemente la memoria y los cores de la GPU? Usa herramientas como `nvidia-smi` para monitorizar la utilización de la GPU (GPU-Util). Si está constantemente baja, podrías estar usando una GPU demasiado potente (y cara).
- **Herramientas de Ayuda:** AWS Compute Optimizer ofrece recomendaciones de right-sizing basadas en ML.

Estrategia 2: Scheduling y Auto Scaling

Scheduling (Programación)

La nube no es como un centro de datos on-premise; no tienes que dejarlo todo encendido 24/7.

Acción: Automatiza el apagado de recursos de desarrollo y pruebas durante las noches y fines de semana.

Herramientas: AWS Instance Scheduler, o simplemente scripts de Lambda con eventos de EventBridge (CloudWatch Events).

Auto Scaling Eficiente

Para los endpoints de inferencia, no aprovisiones para el pico de carga.

Acción: Configura un Auto Scaling Group que ajuste el número de instancias en función de la demanda real (ej: uso de CPU, número de peticiones por segundo).
Beneficio: Solo pagas por la capacidad que necesitas en cada momento.

Estrategia 3: Optimización de Almacenamiento y Datos

Ciclo de Vida del Almacenamiento

No todos los datos necesitan el mismo nivel de acceso.

Acción: Configura políticas de ciclo de vida en S3. Ejemplo: Después de 30 días, mueve los logs de S3 Standard a S3 Infrequent Access (más barato). Después de 90 días, archívalos en S3 Glacier.

Limpieza de Recursos

Los modelos viejos, datasets intermedios y snapshots de EBS olvidados son "basura espacial" que cuesta dinero.

Acción: Crea scripts o procesos para limpiar artefactos de experimentos antiguos.

Optimización de Transferencia de Datos

VPC Endpoints: Para comunicarte con servicios de AWS (como S3) desde tu VPC sin pasar por la internet pública, evitando costes de NAT Gateway y mejorando la seguridad.

Estrategia 4: Optimización a Nivel de Código y Modelo

La optimización de costes también es responsabilidad del Ingeniero de ML.

Eficiencia del Modelo

Un modelo más pequeño y eficiente (ej: usando cuantización, podado o destilación) requiere menos cómputo para la inferencia, lo que permite usar instancias más baratas.

Eficiencia del Código

Asegúrate de que tu código de entrenamiento satura la GPU. Un mal pipeline de datos (lectura lenta desde S3) puede dejar la GPU ociosa, desperdiciando dinero. Usa herramientas como AWS FSx for Lustre para I/O de alto rendimiento.

Consultas Eficientes

Si usas Amazon Athena para analizar datos en tu Data Lake, optimiza tus consultas. Usa formatos de datos columnares (Parquet, ORC) y particiona tus datos para escanear menos información, reduciendo drásticamente los costes de consulta.

Mini-Quiz de Repaso

Pregunta 1

¿Qué modelo de precios ofrece el mayor descuento pero puede ser interrumpido?

- A) On-Demand
- B) Reservado
- C) Spot
- D) Dedicado

Pregunta 2

¿Cuál de estos es típicamente el mayor impulsor de coste en un proyecto de IA?

- A) Almacenamiento
- B) Transferencia de Datos
- C) Cómputo (GPU/CPU)
- D) Logging

Pregunta 3

¿Para qué sirve principalmente el etiquetado (tagging) de recursos?

- A) Mejorar la seguridad
- B) Asignar y rastrear costes
- C) Aumentar el rendimiento
- D) Facilitar el despliegue

Pregunta 4

Mover 1TB de datos desde tu PC a S3 (Ingress) es generalmente más caro que mover 1TB desde S3 a Internet (Egress).

- A) Verdadero
- B) Falso

Construyendo una Cultura de Conciencia de Costes

La tecnología y las herramientas son solo una parte. La optimización sostenible requiere un cambio cultural.

Involucrar a los Ingenieros

Los desarrolladores de IA deben ver los costes como una métrica más de rendimiento de su aplicación, junto con la latencia o la precisión.

Visibilidad

Proporciona a los equipos dashboards sencillos que muestren el coste de los servicios que poseen.

Incentivos

Algunas empresas "gamifican" el ahorro, reconociendo a los equipos que logran las optimizaciones más innovadoras.

Revisión Continua

La optimización de costes debe ser un punto recurrente en las reuniones de equipo. ¿Qué hemos aprendido este mes? ¿Dónde podemos mejorar?

Resumen y Key Takeaways

FinOps es una cultura, no solo un equipo

Combina responsabilidad financiera con agilidad en la nube.

No existe un "mejor" modelo de precios

La estrategia óptima es casi siempre una mezcla de On-Demand, Reservas/Savings Plans y Spot.

El cómputo con GPU es el rey de los costes en IA

Cada hora cuenta.

La visibilidad es el primer paso

Usa Cost Explorer y establece presupuestos. ¡Etiqueta todo!

La optimización es un proceso continuo

Right-sizing, scheduling y limpieza de recursos deben ser hábitos, no eventos puntuales.

¿Preguntas?

Recursos adicionales recomendados:

- Introduction to FinOps - The FinOps Foundation (YouTube)
- AWS re:Invent - Cost optimization for machine learning on AWS