Tecnológico de Monterrey

Maestría en Inteligencia Artificial Aplicada

Proyecto Integrador

Avance 6. Conclusiones clave

Equipo 8

Ignacio Antonio Ruiz Guerra	A00889972
Fernando Ramírez Gómez	A01298109
Juan Pablo Noguerón Morales	A01097897

Análisis de Costos para Modelo de IA en OpenAI y Proveedores de Nube

Resumen Ejecutivo

Este análisis estima los costos asociados con el uso del modelo Llama 3.2, incluyendo tanto los costos de tokens en OpenAI como los recursos de infraestructura necesarios en diversos entornos de nube (AWS, Azure, Google Cloud, IBM Watson). El objetivo es ofrecer una comparación exhaustiva de las diferentes opciones disponibles, proporcionando un marco que permita a los desarrolladores y el cliente tomar decisiones informadas sobre la implementación de soluciones basadas en inteligencia artificial.

---

1. Costos de OpenAI en función de tokens y usuarios

El modelo Llama 3.2, al igual que otros modelos de OpenAI, tiene un costo asociado por el uso de tokens, donde la tarifa estándar aproximada de OpenAI para modelos GPT-4 es de $0.02 por cada 1,000 tokens, que son las unidades de medida para la cantidad de texto procesado. A medida que la longitud de los prompts y las respuestas varían, los costos por token se vuelven críticos para la planificación financiera del proyecto. Por ejemplo, se ha calculado que un usuario promedio podría generar múltiples interacciones al mes, y con una proyección de 30,000 usuarios mensuales, los costos pueden aumentar significativamente. Este análisis no solo toma en cuenta el costo inmediato por token, sino que también considera el costo anual proyectado, lo que resulta en una estimación más precisa para la planificación a largo plazo ursos de Infraestructura en Nubes

Tabla de Costos por Token de OpenAI con Proyección Anual

Tamaño del Prompt (Tokens)	Costo por 1,000 Tokens (USD)	Consultas por Usuario (Mensual)	Costo Mensual (USD)	Costo Anual (USD)
100 Tokens	**$0.02**	**30,000**	**$60**	**$720**
300 Tokens	**$0.02**	**30,000**	**$180**	**$2,160**
500 Tokens	**$0.02**	**30,000**	**$300**	**$3,600**
700 Tokens	**$0.02**	**30,000**	**$420**	**$5,040**
1,000 Tokens	**$0.02**	**30,000**	**$600**	**$7,200**

Explicación del Cálculo de Costos

A pesar que el **costo por cada 1,000 tokens es constante** en $0.02, los **costos totales varían** dependiendo del tamaño del prompt en tokens (de 100 a 1,000 en este caso). Este cambio se debe a que, cuanto mayor es el número de tokens en un prompt, mayor es el costo por consulta:

- **Costo por Consulta**: Se calcula multiplicando el número de tokens de cada prompt por el costo unitario de $0.02 por cada

1,000 tokens. Un prompt de 100 tokens costará $0.002, mientras que un prompt de 1,000 tokens costará $0.02.

- **Costo Total Mensual y Anual**: El costo por consulta se multiplica por el número de usuarios al mes (30,000) y luego se proyecta anualmente. A mayor longitud del prompt, el costo mensual y anual incrementa proporcionalmente debido a la mayor cantidad de tokens procesados.

Ejemplo:

- Para un prompt de 100 tokens: $0.02 * (100 / 1,000) = $0.002 por consulta, resultando en $60 mensuales y $720 anuales.
- Para un prompt de 1,000 tokens: $0.02 * (1,000 / 1,000) = $0.02 por consulta, con un costo mensual de $600 y anual de $7,200.

---

2. Costos de Infraestructura de Nube para Modelos de IA

Para la implementación y entrenamiento del modelo Llama 3.2, revisamos los costos de instancias con GPU en Amazon Web Services (AWS), Microsoft Azure, Google Cloud y IBM Watson. Estas infraestructuras permiten el despliegue y entrenamiento eficiente de modelos grandes, usando principalmente GPUs NVIDIA A100, que son estándar en cargas de trabajo de IA de alto rendimiento.

Además de los costos de tokens, es fundamental evaluar los recursos de infraestructura necesarios para ejecutar el modelo Llama 3.2 en diferentes entornos de nube. Los principales proveedores de nube ofrecen una variedad de opciones de precios basadas en el uso, que incluyen tarifas por procesamiento, almacenamiento y transferencias de datos. Por ejemplo, AWS utiliza un modelo de precios basado en el consumo, donde los

costos varían según el tipo de instancia y los servicios utilizados. Similarmente, Google Cloud y Azure ofrecen modelos de precios competitivos que se ajustan a las necesidades de escalabilidad y rendimiento.

Tablas de Costos de Infraestructura por Proveedor de Servicios en la Nube

1. Amazon Web Services (AWS)

Tipo de Servicio	Costo Mensual (USD)	Costo Anual (USD)	Posibles Upgrades
EC2 (Instancia t3.medium)	**$60**	**$720**	**Cambiar a t3.large ($120/mes)**
S3 (Almacenamiento, 1TB)	**$23**	**$276**	**Escalar a 5TB ($115/mes)**
Lambda (1M solicitudes)	**$0.20**	**$2.40**	**Aumentar a 5M solicitudes ($1.00/mes)**
Total	**$83.20**	**$996.40**	

---

2. Microsoft Azure

Tipo de Servicio	Costo Mensual (USD)	Costo Anual (USD)	Posibles Upgrades
VM (B2s, 2 vCPUs)	$50	$600	Cambiar a B4ms ($100/més)
Blob Storage (1TB)	$20	$240	Escalar a 5TB ($100/mes)
Functions (1M ejecuciones)	$0.25	$3.00	Aumentar a 5 M ejecuciones ($1.25/mes)
Total	$70.25	$843.00	

---

3. Google Cloud Platform (GCP)

Tipo de Servicio	Costo Mensual (USD)	Costo Anual (USD)	Posibles Upgrades
Compute Engine (n1-standard-1)	$50	$600	Cambiar a n1-standard-2 ($100/mes)
Cloud Storage (1TB)	$20	$240	Escalar a 5TB ($100/mes)

Cloud Functions (1M invocaciones)	$0.40	$4.80	Aumentar a 5M invocaciones ($2.00/mes)
Total	$70.40	$844.80	

---

4. IBM Watson

Tipo de Servicio	Costo Mensual (USD)	Costo Anual (USD)	Posibles Upgrades
Watson Assistant (1M mensajes)	$120	$1,440	Aumentar a 5M mensajes ($600/mes)
Almacenamiento (1TB)	$30	$360	Escalar a 5TB ($150/mes)
Otros Servicios (AI, etc.)	$50	$600	
Total	$200	$2,400	

Resumen de Costos por Proveedor

Proveedor	Costo Mensual Total (USD)	Costo Anual Total (USD)
AWS	**$83.20**	**$996.40**
Azure	**$70.25**	**$843.00**
GCP	**$70.40**	**$844.80**
IBM Watson	**$200**	**$2,400**

Consideraciones de Costo

Estos costos son estimaciones basadas en los precios actuales de los proveedores y pueden variar dependiendo de los requisitos específicos del proyecto y del uso real. Cada proveedor también ofrece diferentes niveles de servicio y opciones de escalabilidad, lo que permite adaptar la infraestructura a las necesidades de carga y demanda.

La comparación de precios no solo debe centrarse en los costos directos, sino también en los beneficios adicionales que cada plataforma puede proporcionar. Por ejemplo, algunas nubes ofrecen herramientas de gestión de datos y análisis que pueden facilitar el desarrollo de aplicaciones de IA, mientras que otras pueden presentar ventajas en términos de latencia y soporte técnico. La evaluación de estas variables es esencial para determinar la opción más rentable y eficiente para el uso de

modelos de lenguaje como Llama 3.2

Ventajas y Limitaciones:

- **AWS**: Ofrece la infraestructura de IA más completa, con escalabilidad y servicios avanzados como SageMaker. Sin embargo, es el más costoso en este análisis.
- **Azure**: Brinda capacidades de interconexión avanzada entre GPUs y es competitivo en costos. Sin embargo, puede requerir experiencia adicional en la configuración de sus herramientas de ML.
- **Google Cloud**: Ofrece una de las opciones más económicas y flexibles para la IA distribuida. Se deben gestionar cuidadosamente los costos de salida de datos, que son adicionales.
- **IBM Watson**: Ideal para soluciones empresariales específicas, aunque es menos competitivo en cuanto a variedad de instancias y costo comparado con otras plataformas.

Conclusión

Este análisis proporciona una visión integral de los costos asociados con el uso del modelo Llama 3.2 en un contexto de alta demanda. La evaluación de los costos de tokens en OpenAI y los recursos de infraestructura en diferentes entornos de nube permitiendo anticipar los gastos y optimizar la asignación de recursos. Con una proyección anual para 30,000 usuarios mensuales, es evidente que una planificación cuidadosa y un entendimiento claro de las variables de costos son esenciales para el éxito de cualquier implementación de IA.

Referencias

- Amazon Web Services. (n.d.). Pricing. Retrieved from AWS Pricing
- Microsoft Azure. (n.d.). Pricing Calculator. Retrieved from Azure Pricing
- Google Cloud. (n.d.). *Pricing*. Retrieved from Google Pricing
- IBM. (n.d.). *Watson Pricing*. Retrieved from IBM Pricing