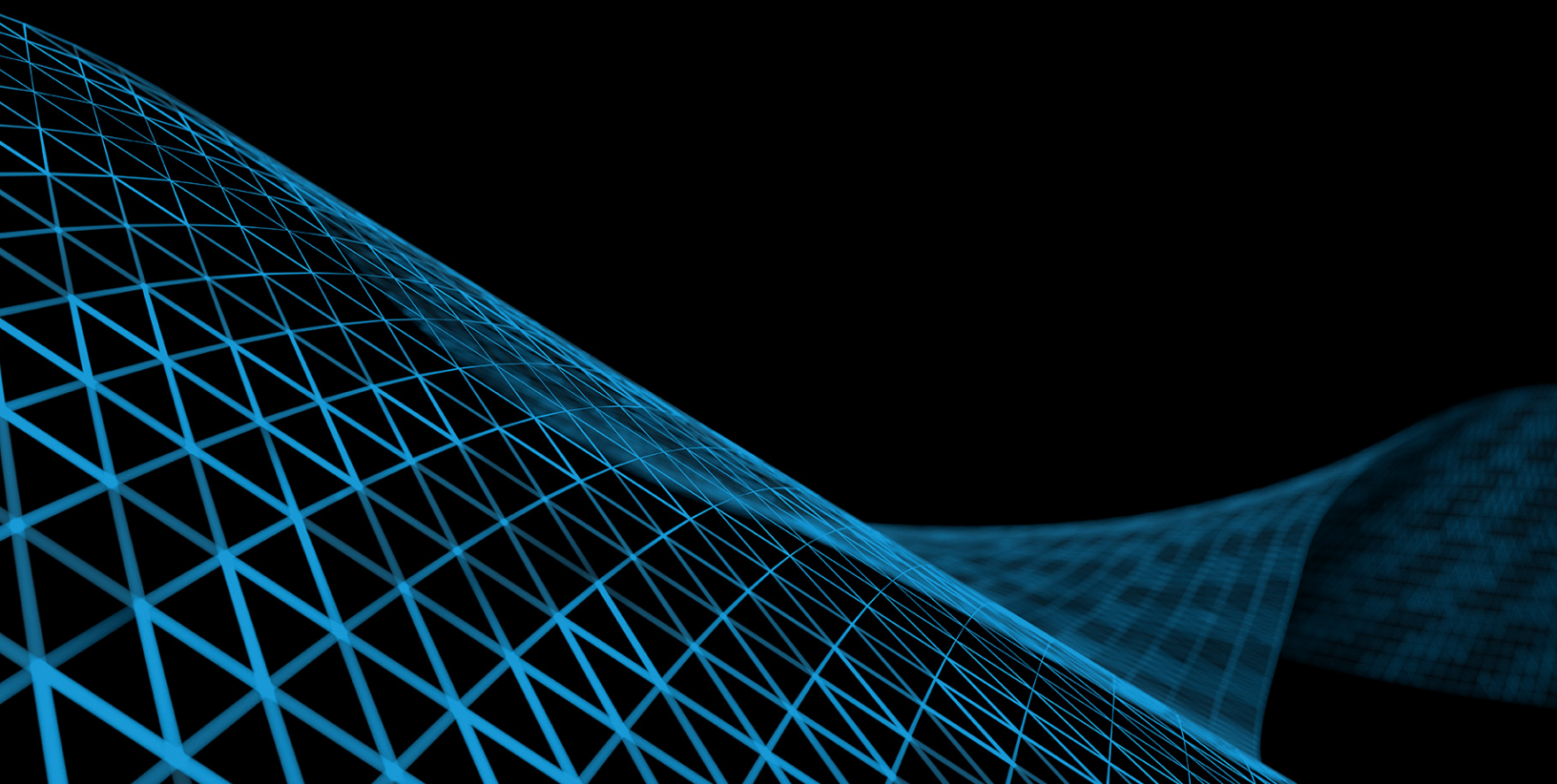
TechTarget | Custom Media
E-Book

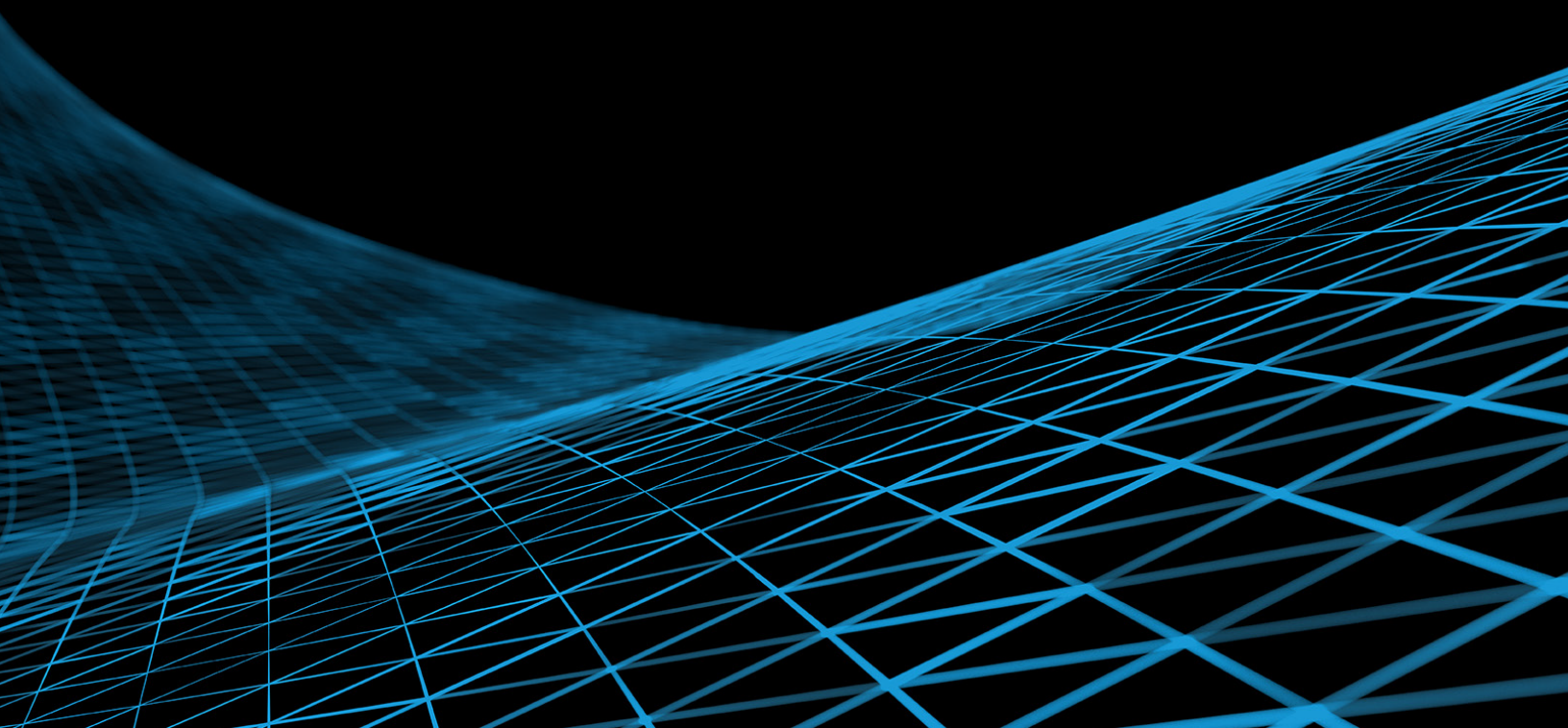
Cómo la IA lo está cambiando todo, desde el centro de datos hasta el Edge y la nube

El impacto transformador de la IA no debe subestimarse. Está cambiando el modo en el que la nube y el edge computing aportan valor a las organizaciones de todos los sectores, en cualquier territorio e independientemente del caso de uso. Este eBook ofrece una perspectiva práctica sobre el papel cada vez más importante de la IA en la nube distribuida.

Akamai



Índice



```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Introducción

Por qué es importante este eBook

Ya sea un responsable de la toma de decisiones de TI, un alto ejecutivo, una parte interesada de la línea de negocio o un miembro de la junta directiva, es probable que esté dedicando cada vez más tiempo a leer, hablar y meditar sobre el impacto de la IA en su organización y, muy posiblemente, en su persona, a nivel particular y profesional.

En su trayectoria de desarrollo en el mercado, la IA avanzó mucho más rápido que la mayoría de las tecnologías más populares que la precedieron. Esta es una de las pocas tecnologías importantes de los últimos años en las que la expectación inicial del mercado estaba plenamente justificada y que ha obtenido resultados tangibles y observables (compartiremos algunas estadísticas que ilustran este importante hecho).

Con todo, este eBook no se limita a analizar el crecimiento y el potencial de la IA en un contexto real. También estudia la IA desde una perspectiva importante: IA cerca de los usuarios y en la nube distribuida. Mucho se ha hablado de la importancia de ubicar los datos cerca de los usuarios y en la nube; ahora, ha comenzado la era de la inferencia de IA descentralizada y la IA en la nube distribuida.

Esto nos lleva de vuelta al título de esta sección: "Por qué es importante este eBook". Ofrecemos cuatro razones por las que los responsables de la toma de decisiones, las personas influyentes y los encargados de la seguridad necesitan leer este eBook:

1. **El edge computing y el cloud computing son las nuevas fronteras de la innovación y la creación de valor en lo relativo a la IA.** Al crear e implementar capacidades de IA fuera del centro de datos, las organizaciones pueden aprovechar las ventajas de nuevos conocimientos, modelos de negocio y soluciones que siguen el mismo camino que el mercado de TI en general ha tomado en la última década.
2. **La IA plantea grandes retos en materia de regulación y gobernanza.** Si el uso y las prácticas de la "IA responsable" se le antojaban difíciles cuando la IA se implementó principalmente en entornos de pruebas locales, considere los problemas de cumplimiento, protección de datos y seguridad que plantea la IA cerca de los usuarios y en la nube distribuida. Su organización debe definir claramente esta área o corre el riesgo de encontrarse con infinidad de problemas.
3. **El impacto de la IA en los patrones y las prácticas cambiantes de la plantilla va a ser incluso mayor de lo esperado.** Si bien no hay duda de que muchas funciones rutinarias de TI pasarán drásticamente de los trabajadores humanos a la IA, trasladar la IA fuera del centro de datos creará oportunidades muy interesantes, incluso inspiradoras, tanto para los trabajadores de TI como para los empleados de otros departamentos.
4. **No hay tiempo que perder.** Las iniciativas de IA están alcanzando los primeros puestos de la lista de prioridades estratégicas de todas las organizaciones. Si su organización no aprovechó al máximo la IA local para probar nuevos casos de uso y aprender más sobre lo que la IA puede hacer, sin duda está rezagada. No obstante, la transición hacia una IA cerca de los usuarios y en la nube distribuida representa una nueva oportunidad.

Además, otra razón por la que confiamos en que este eBook le resultará útil es la siguiente: al final, ofrecemos consejos prácticos para que su organización pueda aprovechar al máximo la transición hacia la IA en la nube distribuida.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

Capítulo 1

La revolución de la IA y cómo ha cambiado ya la dinámica del mercado

Dado que la IA existe desde hace décadas, ¿es justo llamar a la era actual una "revolución de la IA"? ¿O la IA simplemente está evolucionando —a un ritmo acelerado— a partir de sus formas y funciones anteriores? No son preguntas retóricas: nuestro sector tiende a hablar de revoluciones técnicas que aparecen en escena aparentemente de la noche a la mañana, prácticamente sin que nadie las detecte, salvo aquellos expertos técnicos que ya han experimentado con ellas en los laboratorios.

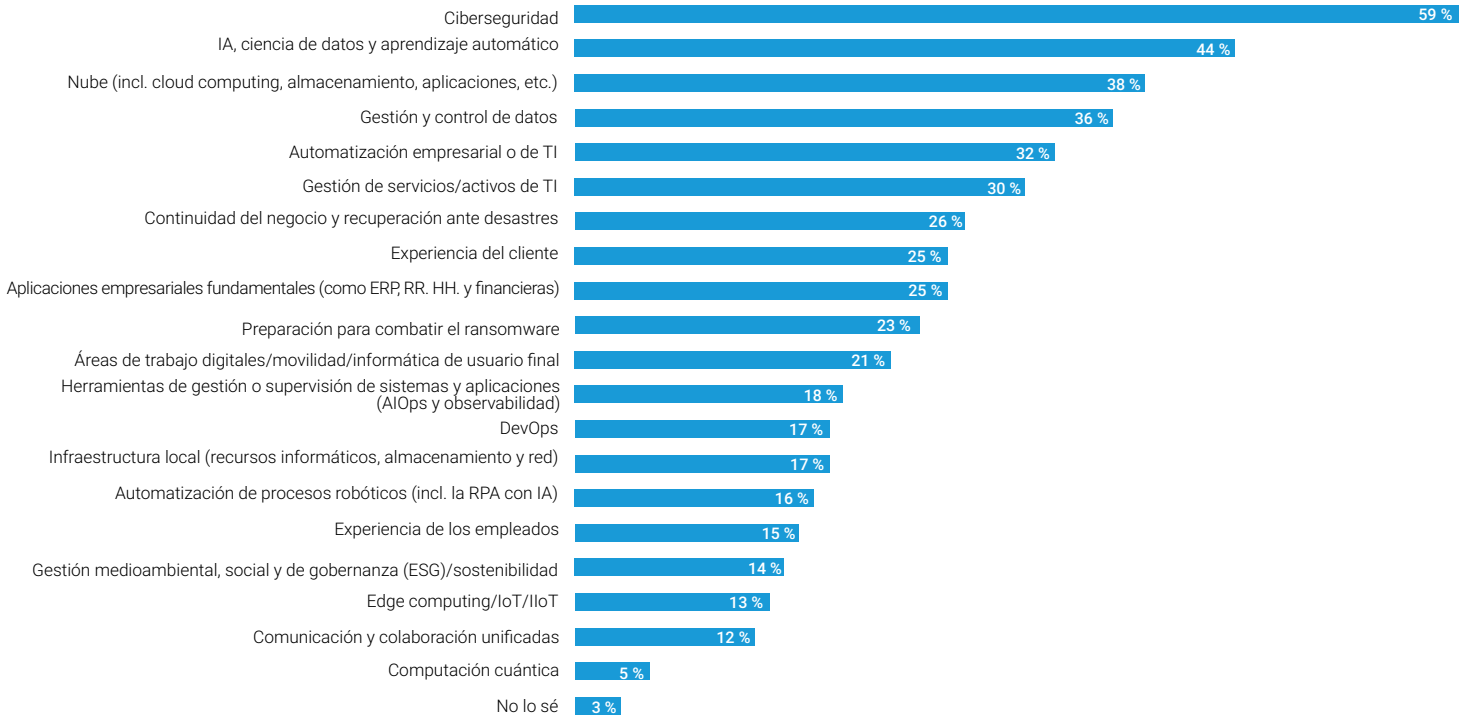
A nuestro juicio, resulta adecuado denominar a esta era como una "revolución de la IA" por varios motivos. En primer lugar, es innegable que el crecimiento del mercado de la IA (y la importancia que los directivos de las organizaciones otorgan ahora a esta tecnología) ha sido impresionante. Sus tasas de crecimiento y el impacto actual en el mercado superan con creces los de otros hitos tecnológicos en el desarrollo del sector de TI, como los ordenadores personales, las redes de área local, el cloud computing y el Internet de las cosas (IoT).

Los siguientes datos ilustran el impacto, la importancia y el potencial de la revolución de la IA:

- El gasto global en hardware, software y servicios relacionados con la IA superó los 600 000 millones de dólares a finales de 2024, y aumentará a más de 2,7 billones de dólares en 2032; esto supone una tasa de crecimiento anual compuesto de más del 20 %.[1]
- El 97 % de los directivos empresariales afirmó que sus organizaciones están obteniendo un retorno positivo de sus inversiones en IA.[2]
- El porcentaje de directores ejecutivos que situaron la IA como una de sus dos prioridades tecnológicas principales aumentó del 4 % al 24 % en solo un año.[3]
- Como se muestra en la figura que aparece a continuación, en los últimos dos años casi la mitad de las empresas de EE. UU. han afirmado que la IA es significativamente más importante para el futuro de su organización; solo la ciberseguridad la supera como iniciativa corporativa.[4]

Figura 1. La seguridad se alza entre la IA, la nube y la gestión de datos como iniciativas tecnológicas cruciales

¿Cuáles de las siguientes iniciativas tecnológicas generales han adquirido una importancia significativamente mayor para el futuro de su organización en los últimos dos años? (Porcentaje de encuestados, N=1351, excl. respuestas múltiples)



1 Fuente: Fortune Business Insights, "Artificial Intelligence market size ... 2024-2032" (Tamaño del mercado de la inteligencia artificial ... 2024-2032), diciembre de 2024.

2 Fuente: Lizzie McWilliams, "Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks" (Las inversiones en inteligencia artificial se mantendrán sólidas en 2025, pero los directivos sénior reconocen riesgos emergentes), EY.com, 10 de diciembre de 2024.

3 Fuente: LinkedIn, la encuesta de CEO de Gartner de 2024 revela que la IA es la principal prioridad estratégica, mayo de 2024.

4 Fuente: Resultados completos de la encuesta de Enterprise Strategy Group, *2025 Technology Spending Intentions Survey* (Encuesta sobre intención de gasto en tecnología en 2025), diciembre de 2024.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


De hecho, el aumento en la adopción de la IA y la inversión de capital correspondiente ha llegado a superar al del cloud computing, uno de los segmentos más vitales y de más rápido crecimiento del panorama de TI. La razón por la que el crecimiento de la IA eclipsa al de la nube es bastante sencilla: mientras que el cloud computing representa una forma potente y altamente aprovechable de desarrollar y prestar servicios de TI, la IA está cambiando radicalmente nuestro modo de trabajar, vivir e interactuar en casi todos los aspectos de nuestras vidas.

Además, la rápida adopción de la IA puso de manifiesto las limitaciones de las arquitecturas de cloud computing tradicionales. Por ejemplo, los modelos de nube centralizados tienen dificultades para satisfacer las exigentes demandas de rendimiento de la IA, que hace un uso intensivo de los datos en tiempo real. En esencia, la IA ha cambiado drásticamente las expectativas en torno a las arquitecturas de nube en lo que respecta a cuestiones como el rendimiento, los costes, las habilidades de las personas, la seguridad y el control de datos.

¿Qué está impulsando el crecimiento estratosférico de la IA y su impacto en el mercado? Un factor importante es el uso mucho más sofisticado y perspicaz que las organizaciones han hecho de los modelos de IA durante el último año. Debido al rápido aumento de la utilidad y la facilidad de uso de los modelos de IA, así como al impresionante retorno de la inversión que han logrado muchas organizaciones, han surgido casos de uso importantes para reforzar la idea inicial sobre cómo utilizar la IA y crear nuevas corrientes de pensamiento en torno a las formas de beneficiarse de los modelos de IA.

Consideremos lo siguiente:

- La **previsión empresarial** siempre ha sido una parte esencial del éxito de las organizaciones, y la IA ha hecho que esa función sea más precisa, oportuna y práctica. La predicción de tendencias futuras con herramientas como los modelos predictivos de IA es incluso más avanzada que hace una década, cuando las organizaciones adoptaron el análisis de Big Data, ya que los modelos actuales de IA están potenciados con datos propios de las organizaciones. Esto permite a los directivos empresariales, además de a los científicos de datos, examinar grandes volúmenes de datos de forma más rápida y precisa que con los modelos tradicionales.
- La **elaboración de resúmenes de documentos** solía ser una labor de los oficinistas, pero la IA generativa (GenAI) ha cambiado las tornas por completo. Los documentos de gran extensión, incluidos aquellos repletos de contenido denso y tecnicismos, se revisan, analizan y sintetizan automáticamente en una fracción del tiempo que solía llevar este trabajo. Esto no solo ofrece una información mejor fundamentada, sino que permite aprovechar el potencial de los profesionales de TI y de otros campos de formas más estratégicas e innovadoras, lo que mejora la productividad y la experiencia de los empleados.
- Actualmente, el **modelado de rotación** es un elemento fundamental en la gestión de las relaciones con los clientes, ya que ayuda a las organizaciones a tomar decisiones mejores y más relevantes según el contexto para predecir cómo, cuándo y por qué esos clientes pueden optar por la competencia, o por qué pueden tener una mayor disposición a comprar.
- Los **chatbots basados en IA** están transformando las funciones de servicio al cliente, gestión de servicios de TI y recursos humanos al posibilitar un autoservicio inteligente en tiempo real para muchas solicitudes rutinarias. Estos chatbots se utilizan incluso para casos de uso de ingeniería y ciencias de la computación altamente técnicos, como la generación de código y el modelado de nuevas funciones.

Estos casos de uso han servido de ayuda para desarrollar muchas otras aplicaciones, flujos de trabajo y casos de uso relacionados con la IA, como la detección de amenazas de ciberseguridad, el análisis de datos o la inteligencia empresarial, el análisis competitivo, la creación rápida de prototipos, la gestión del cumplimiento normativo y mucho más.

Por supuesto, ninguna nueva tecnología, ni siquiera una con tanto impulso y potencial como la IA, está exenta de desafíos en su desarrollo, implementación y gestión. Estos desafíos pueden ser técnicos, como la falta de calidad de los datos, la creación de la infraestructura adecuada o la garantía del elevado rendimiento necesario para impulsar modelos de lenguaje de gran tamaño (LLM). También pueden guardar relación con el negocio o el riesgo, como el control de datos, la garantía del uso responsable de la IA o la superación de la gran y creciente brecha de habilidades en materia de IA.

Sea como fuere, los indicios iniciales son que las organizaciones comprenden esos desafíos y están afrontándolos con los recursos adecuados y el respaldo de los altos ejecutivos y la junta directiva. En el próximo capítulo, explicaremos uno de los desarrollos clave que hacen que la IA sea más que un proyecto científico: cambiar el foco de la IA del entrenamiento de modelos a la inferencia.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

Capítulo 2

El cambio de paradigma de la IA: del entrenamiento a la inferencia

El entrenamiento de modelos de IA fue un factor crítico en el desarrollo del mercado y fue necesariamente el foco del énfasis inicial de las organizaciones en las iniciativas de IA. Las organizaciones no tardaron en concluir que la falta de calidad de los datos (o quizás la falta del tipo y volumen adecuados de datos) afectaba el desarrollo de los modelos de IA. Según una investigación de Enterprise Strategy Group de Informa TechTarget, el 37 % de las organizaciones citaron los problemas de calidad de los datos como su principal desafío al implementar GenAI, lo que lo convierte en el segundo desafío más citado después de la falta de competencias en materia de IA.[5]

Como las organizaciones reconocieron que aprovechar sus propios datos era la mejor manera de resolver los problemas de calidad de los datos y crear LLM más precisos y receptivos, y adaptados al contexto, ahora las organizaciones pueden crear el modelo de IA adecuado con mayor facilidad. Este enfoque suele ser mucho mejor que comprar modelos de IA comerciales estándar, entrenados con datos de terceros, que pueden crear sesgos e imprecisiones en los datos o carecer de capacidades de personalización. Como práctica estándar, la mayoría de las organizaciones centraron sus esfuerzos en materia de entrenamiento de LLM en entornos de nube centralizados, un enfoque eficaz para la fase inicial del desarrollo de LLM.

El entrenamiento de modelos de IA ha madurado y se ha estabilizado en calidad e integridad. Las organizaciones ya están asignando menos recursos a sus modelos de entrenamiento internos, especialmente porque el ajuste y la optimización de la IA reducen la necesidad de una infraestructura centrada en la GPU y con uso intensivo de recursos informáticos.

Mike Leone, director de prácticas de Enterprise Strategy Group para IA, análisis e inteligencia empresarial, ofreció una perspectiva independiente sobre esta transición del entrenamiento a la inferencia:

"A medida que los modelos preentrenados y las aplicaciones de IA en tiempo real maduran y se hacen más prominentes, la inferencia pasa a jugar un papel central en la entrega del valor prometido de la IA de manera eficiente y rentable", asegura. "Este cambio de enfoque está permitiendo implementaciones escalables en todos los sectores, y uno de los principales impulsores de esto son los continuos avances que se están realizando tanto en la pila de software como en el hardware. Los proveedores están sumamente concentrados en dimensionar y optimizar la infraestructura de IA, mejorar la eficiencia del modelo y gestionar los costes operativos corrientes. Por supuesto, a medida que la inferencia democratiza la disponibilidad de la IA para las masas, introduce desafíos como la escalabilidad, las preocupaciones sobre la privacidad y el escrutinio normativo".

La transición del entrenamiento de modelos de IA a aplicaciones de IA basadas en inferencias y casos de uso de alto impacto ya está en marcha. Armand Ruiz, vicepresidente de IBM para plataformas de IA, lo expresó de forma breve y directa: "En mi opinión, la inferencia dominará las cargas de trabajo de IA. … Una vez que un modelo se ha entrenado correctamente, lo ideal es que no haga falta entrenarlo más. Sin embargo, la inferencia es continua".[6]

Cuando se producen transiciones en el mercado, inevitablemente se caracterizan por oscilaciones desproporcionadas en cómo y dónde se invierte el capital. Por ejemplo, una estimación indicó que alrededor del 15 % de los gastos en chips de silicio para IA se destinan al entrenamiento de IA, con un 45 % dedicado a la inferencia de IA en centros de datos y el 40 % restante a la inferencia de IA en el Edge.[7]

5 Fuente: Resultados completos de la encuesta de Enterprise Strategy Group, "*The State of the Generative AI Market: Widespread Transformation Continues*" (Estado del mercado de la IA generativa: la transformación general continúa), septiembre de 2024.

6 Fuente: LinkedIn, Armand Ruiz, diciembre de 2024.

7 Fuente: Jonathan Goldberg, "The AI chip market landscape: Choose your battles carefully" (El panorama del mercado de chips de IA: cómo escoger las batallas sabiamente), TechSpot.com, 30 de mayo de 2023.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Otras investigaciones han señalado una tendencia similar hacia la inferencia, y un estudio de Google destacó esta tendencia de otra manera: Google indicó que el 60 % de su consumo energético centrado en el aprendizaje automático se destina a la inferencia, en comparación con solo el 40 % destinado al entrenamiento.[8]

No es de extrañar que esta transición en la IA del entrenamiento de modelos a la inferencia tenga implicaciones significativas para la infraestructura de IA y las arquitecturas de nube. Esta transformación es especialmente importante para las organizaciones que buscan el entorno y la arquitectura de nube adecuados para el desarrollo, la implementación y la gestión de aplicaciones de IA.

Los modelos de nube centralizados tradicionales siguen funcionando bien para muchas cargas de trabajo empresariales, pero la IA somete a una gran presión a la arquitectura de nube para cumplir con las exigencias de rendimiento de la IA a gran escala. Estos desafíos probablemente se manifestarán como problemas de latencia frustrantes, con un retraso notable entre las consultas y las respuestas correspondientes.

La infraestructura informática también es un aspecto clave que las organizaciones deben tener en cuenta al migrar del entrenamiento de modelos de IA a la inferencia de IA. Las GPU centradas en IA han sido aclamadas por su capacidad para ofrecer la potencia necesaria para entrenar LLM, ya que ofrecen un alto rendimiento computacional y gran ancho de banda de memoria. La inferencia, en cambio, requiere una mentalidad de infraestructura diferente, centrada en la baja latencia y el uso eficiente de los recursos.

En el próximo capítulo, analizaremos en profundidad cómo las organizaciones pueden ayudar a satisfacer las demandas de rendimiento de la inferencia de IA posicionando los recursos de inferencia más cerca del usuario y en una arquitectura de nube distribuida.

8 Fuente: "*The AI boom could use a shocking amount of electricity*" (El auge de la IA podría consumir una cantidad sorprendente de electricidad), Scientific American, 13 de octubre de 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Capítulo 3

El auge de la inferencia de IA descentralizada y la nube distribuida

Como se explicó en el capítulo anterior, la arquitectura tradicional de nube centralizada impone límites y desafíos problemáticos para las organizaciones que buscan optimizar el uso de la IA para crear aplicaciones innovadoras y transformadoras. Se deben considerar y planificar cuestiones como la latencia, las limitaciones del ancho de banda de la red, la garantía del rendimiento a escala y la complejidad de la gestión, a medida que los requisitos de la IA pasan del entrenamiento de modelos a la inferencia.

Las aplicaciones de inferencia de IA modernas y de próxima generación requieren la capacidad de desarrollar, implementar y aprovechar la inferencia donde residen los datos. Cada vez más, esto significa cerca de los usuarios o en una arquitectura de nube distribuida, ágil, escalable y flexible.

Antes de profundizar, vale la pena aclarar ciertas definiciones.

- La *inferencia de IA descentralizada* es un paradigma para crear flujos de trabajo de IA que abarcan centros de datos centralizados (la nube) y dispositivos fuera de la nube, más cercanos a las personas y a los objetos físicos (el Edge). Esto contrasta con la práctica más común, en la que las aplicaciones de IA se desarrollan y ejecutan completamente en la nube, lo que se ha empezado a denominar *IA en la nube*. También difiere de los enfoques de desarrollo de IA más antiguos, en los que se creaban algoritmos de IA en ordenadores de escritorio y luego se implementaban en otros equipos o en hardware especial para realizar tareas como la lectura de números de cheques.

La distribución de servicios en la nube puede ayudar a las organizaciones a cumplir los objetivos de rendimiento de las aplicaciones y tiempo de respuesta, edge computing, requisitos normativos u otras demandas...

- Akamai, un proveedor líder de plataformas para la nube, la seguridad y la distribución de contenido, define la *nube distribuida* como "una forma de cloud computing en la que una empresa utiliza una infraestructura de nube pública repartida por varias ubicaciones geográficas, mientras que las operaciones, el control y las actualizaciones se gestionan de forma centralizada por un único proveedor de servicios de nube pública. La distribución de servicios en la nube puede ayudar a las organizaciones a cumplir los objetivos de rendimiento de las aplicaciones y tiempo de respuesta, edge computing, requisitos normativos u otras demandas que se pueden satisfacer proporcionando servicios en la nube desde ubicaciones más cercanas a los usuarios finales y los dispositivos. El uso de cloud computing distribuido está impulsado por el aumento de las redes del Internet de las cosas (IoT), la inteligencia artificial y otros casos de uso que deben procesar grandes cantidades de datos en tiempo real".

Tanto la inferencia de IA descentralizada como la nube distribuida son vitales para el desarrollo y la utilización de la inferencia de IA, ya que representan formas modernas, eficientes, seguras y fiables de superar los desafíos mencionados anteriormente: alto rendimiento; rendimiento a escala; uso eficiente de los recursos informáticos, de almacenamiento y de red, así como la superación de los problemas de latencia asociados a grandes conjuntos de datos que suelen incluir enormes cantidades de datos no estructurados.

¿Cómo funciona esto en la vida real? Una empresa quería hacer de la personalización un sello distintivo de la experiencia del cliente que ofrecían y necesitaba asegurarse de poder capturar datos de todo el espectro de dispositivos y fuentes posibles. Decidieron utilizar un chatbot de IA basado en LLM que distribuía datos localizados a ubicaciones cercanas a los usuarios. De esta forma, lograron reducir la latencia y los tiempos de respuesta, y esto se tradujo en una mejora enorme del rendimiento y una atención al cliente más rápida y personalizada.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

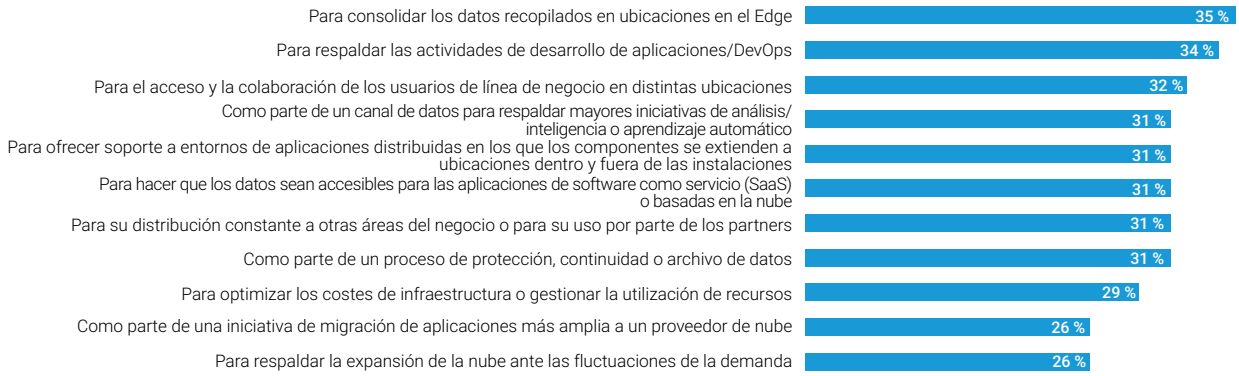

Este enfoque es cada vez más común en las organizaciones que buscan aprovechar las crecientes oportunidades que ofrecen las aplicaciones de inferencia de IA. Para resolver los problemas de rendimiento derivados de las limitaciones de la infraestructura informática y los problemas de latencia, las cargas de trabajo de IA están acercando a un mayor número de usuarios finales a las fuentes de datos para que las utilicen mediante inferencia de IA descentralizada y arquitecturas de nube distribuida.

Por ejemplo, este modelo se utiliza ampliamente en sectores con instalaciones altamente distribuidas y geográficamente dispersas, como el retail, las telecomunicaciones y los medios de comunicación. Este enfoque se está convirtiendo cada vez más en el modelo arquitectónico estándar para la inferencia de IA debido a su capacidad para superar los desafíos de latencia y rendimiento a escala asociados con los complejos requisitos de inferencia de IA.

Un factor clave para la transición a un modelo de nube distribuida para la inferencia de IA es la creciente necesidad de las organizaciones de utilizar el Edge para transferir datos entre centros de datos y proveedores de servicios en la nube (CSP). Enterprise Strategy Group descubrió que el 35 % de las organizaciones que transfieren datos regularmente entre centros de datos y la infraestructura de los CSP lo hacen para consolidar los datos recopilados en ubicaciones en el Edge, lo que lo convierte en la principal motivación para ese movimiento de datos.[9]

Figura 2. El Edge es el principal impulsor de la transferencia de datos entre centros de datos y servicios de nube pública

Ha indicado que su organización transfiere datos regularmente entre sus centros de datos y los servicios de nube pública. ¿Para qué fin(es) mueve su organización los datos entre sus centros de datos y los servicios de nube pública? (Porcentaje de encuestados, N=170, múlt.)



9 Fuente: Informe de investigación de Enterprise Strategy Group, *Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud* (Serie de nubes distribuidas: el estado de la modernización de la infraestructura en la nube distribuida), noviembre de 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh
```

Capítulo 4

Por qué la inferencia de IA descentralizada es el futuro

Desarrollar e implementar la inferencia de IA con datos de ubicaciones en el Edge es clave para ayudar a las organizaciones a aprovechar al máximo la inferencia. Con tantos datos generados y alojados en diversos dispositivos en el Edge —como smartphones, dispositivos portátiles diseñados específicamente, carteles digitales y dispositivos IoT—, las organizaciones deben adoptar una "mentalidad del Edge" para lograr el procesamiento de datos en tiempo real, esencial para la inferencia de IA.

El cambio hacia una inferencia de IA descentralizada para cumplir con los requisitos de este modelo ha sido rápido y profundo. Las investigaciones indican que el gasto en sistemas de inferencia cerca de los usuarios se multiplicará por diez, pasando de 27 000 millones de dólares en 2024 a casi 270 000 millones de dólares en 2032.[10]

La inferencia de IA descentralizada se caracteriza por una serie de consideraciones importantes para ayudar a las organizaciones a desarrollar e implementar aplicaciones de inferencia de IA, como el procesamiento de IA nativo del dispositivo, marcos de seguridad integrados optimizados para dispositivos en el Edge, bajo consumo de energía y baja latencia. Estas características se traducen en un rendimiento más rápido a escala, una comprensión de los datos más precisa y profunda, y una menor inversión en ancho de banda. Esto, a su vez, permite a los miembros de una organización geográficamente dispersos utilizar la inferencia de IA para tomar decisiones más inteligentes y en tiempo real, mucho más cerca de donde se crean, almacenan, procesan y consultan los datos.

Bloomfield es un ejemplo de una organización con visión de futuro que utilizó la inferencia de IA descentralizada para mejorar las operaciones esenciales y simplificar la gestión de datos. La empresa necesitaba gestionar de forma eficiente, rentable y segura grandes volúmenes de datos procedentes de sus cámaras inteligentes distribuidas para informar a los agricultores, que a menudo operan con acceso limitado a la red. Mediante un modelo arquitectónico de inferencia de IA descentralizada, Bloomfield aprovechó la inferencia de IA para determinar la mejor manera de organizar, almacenar y presentar los datos a los agricultores. Esto resultó en una gestión simplificada del estado de los cultivos, un mejor rendimiento de la producción y una toma de decisiones más rápida y precisa, lejos de un centro de datos local.

La capacidad de la inferencia de IA descentralizada para respaldar aplicaciones de inferencia de IA en ubicaciones diversas y distribuidas la hace ideal para casos de uso como los siguientes:

- **Medicina de cabecera y aplicaciones de salud remota**, donde los profesionales sanitarios toman decisiones desde cualquier lugar mediante smartphones, dispositivos IoT y otros dispositivos portátiles ligeros pero potentes. La inferencia de IA descentralizada también es ideal para tomar decisiones en tiempo real utilizando dispositivos inteligentes en el ámbito sanitario, como bombas de infusión, marcapasos, monitores cardíacos y análisis de medicamentos.
- **Ciudades inteligentes**, donde los sistemas en el Edge lo hacen todo, desde controlar el tráfico rodado y supervisar el comportamiento de los sistemas de servicios públicos hasta agilizar el registro de votantes y mejorar la seguridad pública.
- **Vehículos sin conductor**, que utilizan cámaras inteligentes y datos de sensores integrados para evaluar las opciones de ruta, supervisar y mejorar el consumo de combustible, documentar los cambios en el GPS y comunicarse con las autoridades y los organismos de seguridad pública.
- **Retail**, que admite una amplia gama de aplicaciones de inferencia de IA, desde videovigilancia IP y gestión de inventario hasta prevención de robos y seguimiento de los patrones de tráfico de los clientes en tiendas.

A medida que las empresas buscan maneras de reducir la latencia en las respuestas a consultas en tiempo real, la inferencia de IA descentralizada se convertirá en un requisito fundamental para una toma de decisiones empresarial más eficiente, mejor informada y precisa. Las aplicaciones de IA ya no tendrán que enviar ni solicitar datos a centros de datos remotos; en su lugar, procesarán y compartirán datos en una nube distribuida globalmente.

El director de ingeniería de una importante empresa de tecnología publicitaria señaló: "Dedicamos mucho tiempo a optimizar nuestro modelo de IA. Ahora, nos damos cuenta de que la inferencia no debe dejarse para el último momento".

Las organizaciones ya no pueden permitirse el lujo de decidir si deben adoptar la inferencia de IA como un componente principal de su estrategia empresarial. Deben adoptar un mayor sentido de urgencia y utilizar la inferencia de IA descentralizada para inferir, ya que muchos de sus competidores ya están dando grandes pasos hacia el uso de esta tecnología cerca de sus usuarios.

10 Fuente: Fortune Business Insights, "Edge AI Market Size … 2024-2032" (Tamaño del mercado de la IA en el Edge … 2024-2032), diciembre de 2024.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Capítulo 5

Una visión estratégica de la IA, el Edge y la nube distribuida: rumbo futuro e implicaciones

Evaluar, seleccionar y colaborar con un partner tecnológico de confianza es fundamental para el éxito de una iniciativa de inferencia de IA en una organización, pero lograrlo es complejo.

Es fundamental debido a un hecho innegable: la importante brecha en habilidades de IA en cuanto a infraestructura, arquitectura en la nube, casos de uso de inferencia, optimización de aplicaciones e implementación impecable.

Además, también es complejo debido a la falta de candidatos idóneos. Como se mencionó anteriormente, un proyecto exitoso de inferencia de IA debe combinar a la perfección conocimientos técnicos, experiencia en el dominio, comprensión de la seguridad, la gobernanza y la gestión de riesgos, y una atención minuciosa al detalle en la implementación y la gestión continua.

Para desarrollar e implementar la inferencia de IA en toda la empresa física y virtual, desde el núcleo hasta el Edge, la nube y viceversa, las organizaciones necesitan alinearse con un proveedor con experiencia previa. Contar con un partner que comprenda los aspectos técnicos y comerciales fundamentales de un proyecto de IA es indispensable.

Akamai, proveedor líder de soluciones sofisticadas para cloud computing, seguridad y distribución de contenido, cuenta con una sólida trayectoria en el desarrollo, la implementación y la gestión de iniciativas de IA, desde el entrenamiento de modelos hasta la inferencia. Su experiencia práctica con cargas de trabajo complejas y de alto consumo de recursos la prepara para trabajar con una amplia gama de clientes en diferentes casos de uso, sectores y geografías.

Para ayudar a las organizaciones a colmar sin problemas sus necesidades de inferencia de IA, Akamai ha implementado con éxito, de manera sistemática, fiable y segura, una arquitectura de nube distribuida que permite a las organizaciones procesar las cargas de trabajo de IA más cerca del usuario final. Esto ha sido fundamental para reducir la latencia y garantizar un alto rendimiento a escala, ambos esenciales para cargas de trabajo de inferencia de IA con un uso intensivo de recursos.

Una de las principales ventajas de la arquitectura de nube distribuida de Akamai es la capacidad de los desarrolladores para centrarse plenamente en el desarrollo independiente de la plataforma en un marco nativo de la nube basado en estándares abiertos.

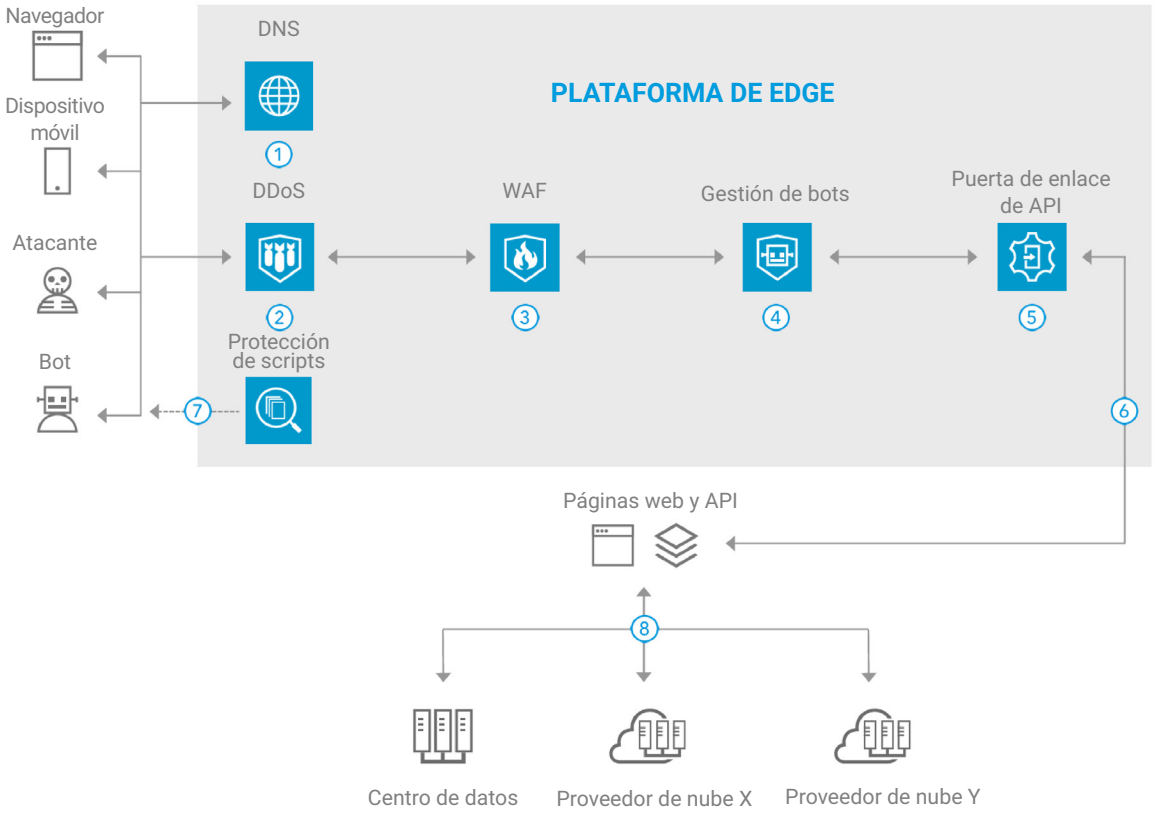
Para lograr un alto rendimiento a escala, Akamai utiliza diversas técnicas de almacenamiento en caché, integradas con software de código abierto, para garantizar un rendimiento suficiente y limitar la latencia. Por ejemplo, muchos clientes de Akamai que utilizan su arquitectura de nube distribuida afirmaron haber logrado una reducción del 50 % en los tiempos de respuesta de los chatbots de atención al cliente, un caso práctico de inferencia muy importante y popular. Este tipo de respuesta casi en tiempo real permite a organizaciones de una amplia gama de sectores, como el retail, la sanidad, los servicios financieros y la alta tecnología, brindar un servicio al cliente más rápido y adaptado a sus necesidades específicas.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Akamai también ofrece una arquitectura de referencia de IA (véase la Figura 3 a continuación) que permite a las organizaciones comenzar a operar rápidamente y obtener un alto valor económico y operativo. Esta arquitectura de referencia es integral, ya que, mediante un conjunto conocido de navegadores y dispositivos definidos por el usuario, las organizaciones pueden implementar una solución de IA empresarial desde el centro de datos principal hasta el Edge y la nube, y viceversa. Esto beneficia a todos los actores de la empresa: desarrolladores, especialistas en TI, administradores de infraestructura, arquitectos de la nube y partes interesadas de la línea de negocio.

Figura 3.

Habida cuenta de los rápidos avances tecnológicos y la creciente complejidad y valor de las cargas de trabajo de IA, las organizaciones deben actuar con rapidez y decisión para superar a la competencia. Es esencial aprovechar al máximo el potencial de la inferencia de IA mediante la adopción de una inferencia de IA descentralizada y una arquitectura de nube distribuida, especialmente cuando la planificación, el desarrollo, la implementación y la gestión corren a cargo de un líder consolidado como Akamai.



```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Conclusión

Qué debería hacer su organización a continuación

Se dice que la IA ha cambiado las reglas del juego para la excelencia de TI y la transformación empresarial. La clave para optimizar el valor de las inversiones en IA de una organización es pasar rápidamente del entrenamiento de modelos de IA a la inferencia de IA.

Para aprovechar al máximo todo lo que ofrece la inferencia de IA, las organizaciones deben desarrollar e implementar sus aplicaciones y cargas de trabajo de inferencia más cerca de donde se realiza el trabajo y donde el personal clave puede lograr resultados extraordinarios con los sistemas. Esto implica pasar de un modelo de cloud computing centralizado a un modelo de nube distribuida, fuertemente influenciado por la IA en el Edge.

Como se ha señalado en otras partes de este eBook, existen muchos problemas que las organizaciones deben considerar y una serie de desafíos que comprender y superar. Esto puede parecer abrumador para las organizaciones que aún no han convertido las iniciativas de IA en una prioridad estratégica, pero hay pasos importantes que pueden y deben dar para avanzar y mantenerse a la vanguardia.

"Las organizaciones que intentan ponerse al día con la IA no deben caer presa del miedo a perderse oportunidades al perseguir la tecnología", subrayó Leone, de Enterprise Strategy Group. "El enfoque debe centrarse en sentar las bases adecuadas mediante medidas como la identificación de problemas empresariales específicos que la IA pueda resolver, la evaluación de su infraestructura de datos y habilidades internas existentes, y la identificación de partners clave que puedan cubrir las carencias. Deben ser estratégicos al identificar uno o dos proyectos piloto de alto impacto y casos de uso que puedan demostrar un valor claro utilizando datos de calidad existentes".

Saber qué medir y cómo medirlo es fundamental para tener éxito en cualquier iniciativa de inferencia de IA.

"Y mientras todo esto sucede, deberían estar desarrollando simultáneamente el talento interno y estableciendo marcos de gobernanza para garantizar un uso responsable. La clave es evitar una implementación apresurada y priorizar medidas estratégicas que permitan construir capacidades fundamentales e, idealmente, duraderas".

Para ello, aquí hay cuatro aspectos que usted y sus homólogos deben comprender y comprometerse a cumplir cuanto antes:

1. **Prestar atención a la selección de casos de uso.** Este es un primer paso esencial, ya que necesitará generar confianza en su capacidad para implementar correctamente la inferencia de IA. También es importante generar confianza entre las partes interesadas de su empresa y los responsables de la aprobación financiera con algunos logros iniciales.
2. **La elección de la infraestructura es importante.** Obviamente, dado que la inferencia de IA requiere un alto consumo de recursos computacionales, la decisión más sencilla es optar por la GPU líder del mercado para el núcleo de su sistema. Pero existen muchas opciones excelentes de GPU y otras infraestructuras de IA. Tómese el tiempo para hacerlo bien (o, mejor aún, confíe en la experiencia y los conocimientos de un partner con experiencia previa).
3. **Conseguir la aprobación de la dirección.** Céntrese en las dos preocupaciones principales de cada ejecutivo y miembro de la junta directiva: ¿cuál es nuestra oportunidad y cuál el riesgo potencial? Asegúrese de que su proyecto y casos de uso se alineen con los objetivos empresariales clave, como mejorar la experiencia del cliente, identificar y aprovechar las oportunidades de mercado con mayor rapidez y crear un proceso de pruebas beta más preciso. Mientras lo hace, identifique y reclute patrocinadores de alto nivel de grupos como TI, la línea de negocio y los miembros de la junta directiva.
4. **Definir y lograr el éxito.** Saber qué medir y cómo medirlo es fundamental para tener éxito en cualquier iniciativa de inferencia de IA. Seleccione métricas (o colabore con su partner para desarrollar métricas personalizadas) que sean realistas, relevantes y lo suficientemente robustas como para generar "grandes logros", especialmente después de identificar algunas oportunidades idóneas con los primeros casos de uso de inferencia de IA.

TechTarget Inc. elaboró el contenido de este documento por encargo de Akamai.