CLOUDFLARE

LIBRO ELECTRÓNICO

5 formas en que el filtrado de DNS puede ayudar a tu estrategia de seguridad de IA

Tabla de contenido

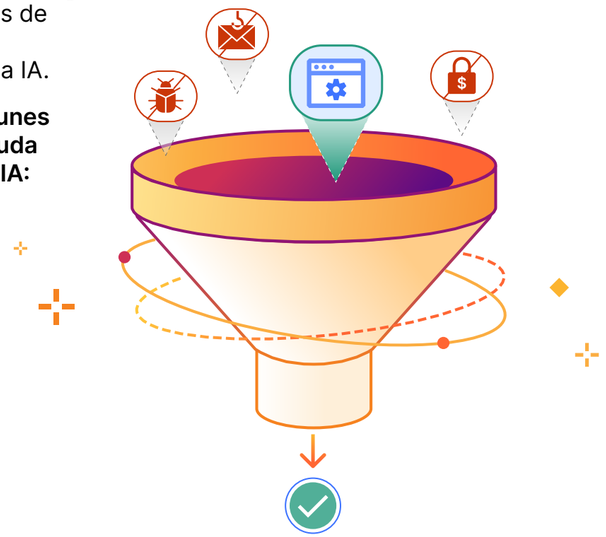
Introducción

El filtrado de DNS ofrece un rápido retorno de la inversión para tu seguridad basada en la IA

A medida que las organizaciones se apresuran a integrar la inteligencia artificial (IA) en sus flujos de trabajo, el interés por aumentar la productividad a menudo oculta fugas de seguridad cada vez mayores. El uso incontrolado de herramientas de IA generativa como ChatGPT o Claude crea una frontera digital al margen de la ley, donde los datos confidenciales corren peligro y la conformidad normativa queda relegada a un segundo plano. Al mismo tiempo, los ciberdelincuentes están utilizando la IA como arma para intensificar sus ataques y explotar esta superficie de ataque en expansión.

Afortunadamente, una de las tecnologías de seguridad más consolidadas, el **filtrado de DNS**, puede ayudar a las empresas a adoptar rápidamente una forma más proactiva y sencilla de mitigar estos riesgos.

El filtrado de DNS (que restringe el contenido web en función de los dominios y las direcciones IP) se considera tradicionalmente una capa de protección sencilla y eficaz para bloquear el malware de Internet y aplicar políticas de uso aceptable. Sin embargo, también es cada vez más habitual que los equipos de informática y seguridad den un primer paso para modernizar su enfoque general de seguridad de la IA.



Este libro electrónico **destaca cinco formas comunes en las que el filtrado de DNS de Cloudflare te ayuda a adaptar tu enfoque de seguridad a la era de la IA:**

1. Detecta la Shadow AI
2. Controla el acceso a la IA
3. Frena las ciberamenazas basadas en la IA
4. Evita la exposición / exfiltración de datos
5. Protege el desarrollo de la IA

Partiendo de esta base inicial, las organizaciones suelen mejorar su visibilidad y sus controles en más entornos, lo que amplía capacidades como las inspecciones HTTP a través de una puerta de enlace web segura (SWG) o una plataforma de perímetro de servicio de acceso seguro (SASE) más completa. Este libro electrónico también explica cómo las organizaciones pueden implementar capacidades SWG y SASE para mejorar aún más su enfoque de seguridad de IA:

Fase de implementación con Cloudflare	Capacidad de ejemplo
Paso 1: Implementa el filtrado de DNS	Análisis del uso de la Shadow AI e implementación de controles de acceso basados en dominios y direcciones IP
Paso 2: Mejora las inspecciones de los SWG	Bloquea las instrucciones de los usuarios en las herramientas de IA basadas en detecciones de datos confidenciales y controles de temas
Paso 3: Amplía la plataforma SASE	Controla el uso de la IA en las comunicaciones entre humanos y la IA y entre máquinas (agéntica)

1 Detecta la Shadow AI / IT

Filtra las consultas DNS para obtener visibilidad básica

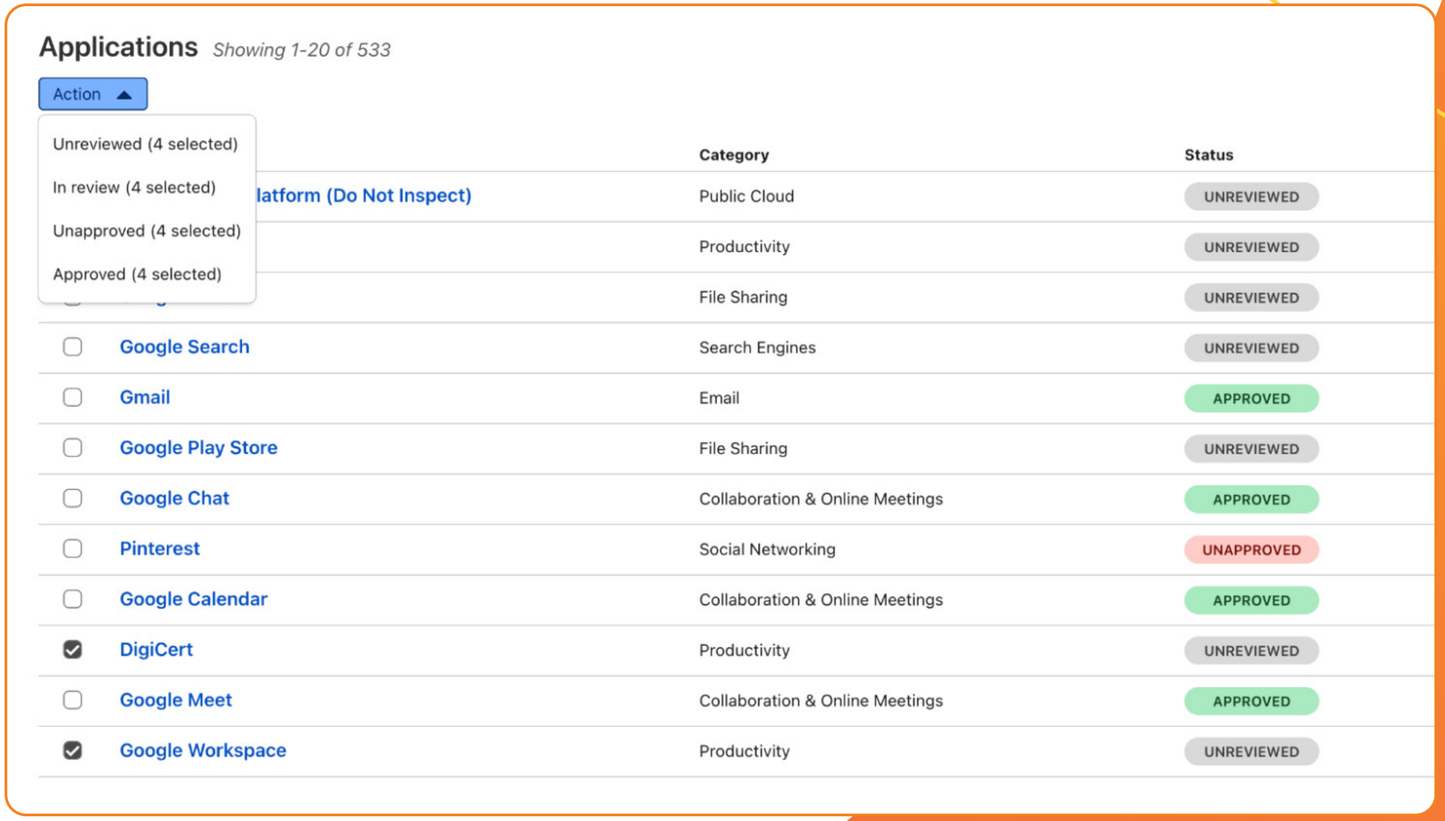
Las organizaciones llevan años lidiando con el uso no autorizado o no aprobado de herramientas SaaS, pero la explosión de las herramientas de IA y la urgencia por utilizarlas está provocando la actual emergencia de la Shadow AI:

El 20 % de las organizaciones sufrió una fuga de seguridad debido a incidentes con la Shadow AI en 2025.[1]

El 85 % de los responsables informáticos afirma que los empleados están adoptando herramientas de IA antes de que el equipo informático pueda evaluarlas.[2]

El filtrado de consultas DNS te ayuda a recuperar la visibilidad básica de los elementos de Shadow AI mediante el seguimiento de todas las consultas DNS realizadas por tus usuarios. Esto te permite:

- **Identificar las aplicaciones** en función de la resolución del dominio (p. ej. chatgpt.com o claude.ai)
- Clasificar y revisar el **estado de aprobación de las aplicaciones** según el dominio (p. ej. aprobado, no aprobado, sin revisar o en revisión). Ver ejemplo a la derecha.
- Evaluar la fiabilidad de una aplicación en función de las **puntuaciones de confianza de la aplicación**. Esta puntuación evalúa no solo los riesgos generales que plantean las herramientas SaaS, como las certificaciones de conformidad y las prácticas de gestión de datos, sino también los riesgos específicos de la IA, como si los datos del usuario se utilizan para el entrenamiento del modelo o si el modelo tiene una tarjeta del sistema publicada que detalla las pruebas de sesgo.



Revisa y marca los estados de las solicitudes en el panel de control

1 Detecta la Shadow AI / IT

Más allá del filtrado de DNS

Mejora los controles de acceso con políticas HTTP

Si bien el filtrado DNS proporciona una referencia básica sobre **quién accede a qué aplicación**, la activación de la inspección HTTP permite obtener una visión más detallada de **lo que hacen dentro de esa aplicación**. Esta visibilidad incluye hasta registros de las instrucciones y respuestas entre los usuarios y las herramientas de IA generativa.

Paneles de control como el siguiente ofrecen análisis agregados de las tendencias a lo largo del tiempo.

Para realizar un análisis más detallado, haz clic en cualquier aplicación de IA para ver los usuarios o grupos específicos que acceden a ella, su frecuencia de uso, ubicación y más detalles.

Panel de control de análisis de Shadow IT

Una línea de investigación habitual es comprender los patrones de transferencia de datos dentro y fuera de las aplicaciones de IA. **A continuación se muestra un ejemplo de análisis de datos descargados / cargados por nombre de host**, que se puede filtrar aún más por usuario, categoría de contenido y otros criterios.

Nombre de host por datos descargados

oneclient.sfx.ms	40,35 MB
www.bing.com	3,17 MB
chatgpt.com	3,14 MB
www.gstatic.com	2,17 MB
gemini.google.com	185,21 KB

Nombre de host por datos cargados

gemini.google.com	2,39 MB
play.google.com	399,97 KB
clients4.google.com	110,00 KB
go.microsoft.com	89,68 KB
www.bing.com	52,67 KB

2 Controla el acceso a la IA

Establece reglas de acceso básicas basadas en categorías de dominios

El filtrado de DNS es una forma sencilla y ágil de evitar que los usuarios accedan a contenidos maliciosos o no deseados en Internet. Para proteger a sus empleados, las organizaciones suelen bloquear todos los dominios y direcciones IP clasificados automáticamente como **riesgos de seguridad** como el malware, el phishing, los servidores de comando y control (C2), las botnets y los destinos de tunelización DNS. También bloquearán **categorías de contenido** como contenido para adultos, apuestas o transmisión de vídeo, así como **aplicaciones clasificadas específicas.** Este filtrado de contenido se utiliza a menudo para aplicar políticas de uso aceptable para empleados o invitados en espacios compartidos como una tienda, un hotel, un hospital o una escuela.

Utiliza categorías de dominio y selectores de aplicaciones para controlar a qué herramientas de IA pueden acceder los usuarios. Por ejemplo, combina dos políticas **para bloquear todas las aplicaciones de IA excepto una aplicación aprobada, ChatGPT:**

Paso 1 **Configura la regla PERMITIR para ChatGPT**

Ver selector de ejemplos

Selector (Obligatorio)
Aplicación

Operador (Obligatorio)
in

Valor
ChatGPT

Paso 2 **Configura la regla BLOQUEAR para todas las demás herramientas de IA**

Ver selector de ejemplos

Selector (Obligatorio)
Categorías de contenido

Operador (Obligatorio)
in

Valor
Inteligencia artificial

Las **acciones de anulación de DNS** incluso permiten que las políticas redirijan el tráfico destinado a dominios de riesgo a recursos internos específicos o servidores "sinkhole" basados en direcciones IP. Por ejemplo, con Cloudflare:

Selector	Operador	Valor	Acción	Invalidar
Nombre de host	IS	www.riskyAI.com	Invalidar	1.2.3.4 (página de política interna de IA)

2 Controla el acceso a la IA *(continuación)*

Más allá del filtrado de DNS

Mejora los controles de acceso con políticas HTTP

Si activan la inspección SWG de proxy completa, las organizaciones pueden habilitar controles de acceso más precisos y flexibles con políticas HTTP. Algunos enfoques populares incluyen:

- **Aplicar políticas para los elementos de Shadow AI basadas en el estado de aprobación de las aplicaciones:** personaliza las reglas para las aplicaciones aprobadas / no aprobadas / sin revisar / en revisión. Bloquear todas las aplicaciones de IA no aprobadas es una opción directa, pero también puedes aplicar acciones más variadas, como las que se indican a continuación.
- **Redirigir el tráfico a URL específicas:** por ejemplo, enviar las solicitudes de los usuarios desde herramientas de IA no aprobadas a una aprobada o a una página de destino educativa.
- **Aislar la sesión en un navegador remoto:** enruta el tráfico para aplicaciones no revisadas, en revisión u otras aplicaciones específicas en un navegador aislado, donde todo el código web se ejecuta en la red de Cloudflare en lugar de en un dispositivo local. El aislamiento te ayuda a proteger los datos controlando las acciones de los usuarios, incluida la restricción de copiar y pegar, las cargas / descargas de archivos, las entradas de teclado y mucho más.
- **Mostrar notificaciones personalizadas a través del cliente de dispositivo:** muestra un mensaje personalizado a través del cliente de dispositivo de Cloudflare cuando se bloquea el tráfico de un usuario. Esto se suele utilizar para explicar la lógica de la decisión de bloqueo.

Si bien estas son políticas de acceso comunes, las políticas HTTP son necesarias para una protección de datos más granular, incluidas las detecciones de prevención de pérdida de datos (DLP), que se analizan en la siguiente sección.

3 Evita las ciberamenazas basadas en la IA

4 Evita la exposición / exfiltración de datos

El filtrado de DNS sigue siendo eficaz contra las amenazas emergentes basadas en la IA y el robo de datos

Los ciberdelincuentes utilizan cada vez más la IA para ejecutar, automatizar y escalar sus ataques, a menudo con el objetivo clásico de extraer datos confidenciales. Estas campañas pueden ser más rápidas, más eficaces y más difíciles de detectar:

El **76 %** de las organizaciones admite que tienen dificultades para igualar la velocidad y la sofisticación de los ataques basados en la IA.[3]

Los investigadores han informado de campañas en las que la IA realiza **el 80-90 %** de un ataque, con solo una mínima intervención humana.[4]

Aunque los titulares suelen centrarse en técnicas novedosas de la IA, como los deepfakes y el malware polimórfico, los atacantes siguen recurriendo a métodos e infraestructuras tradicionales. El filtrado de DNS ofrece una primera línea de defensa eficaz contra ambos extremos de ese abanico.

La tabla de la derecha refleja las amenazas comunes que los servicios de filtrado de DNS bloquean automáticamente y cómo impulsan los ataques de IA para robar datos. En particular, una resolución DNS autoritativa y recursiva como Cloudflare con visibilidad en tiempo real en toda la infraestructura global de Internet (más de 5,7 billones de consultas DNS al día) tiene una telemetría única para impulsar el modelo de detección de amenazas a fin de identificar amenazas, a menudo utilizando la IA y el aprendizaje automático para hacerlo. De esta manera, la seguridad puede utilizar proactivamente la IA para defenderse de la IA.

Amenaza	Función en las campañas basadas en la IA	Cómo ayuda el filtrado de DNS
Dominios de phishing	La IA puede generar señuelos hiperpersonalizados para atraer a los objetivos a dominios de phishing, que a menudo se basan en dominios "similares" (por ejemplo, mybank-security.com en lugar de mybank.com). Allí, los atacantes pueden recopilar credenciales, robar cookies de sesión y cosas peores.	Incluso si un empleado hace clic en un enlace de phishing, la solicitud falla antes de que se pueda cargar la página de phishing.
Devoluciones de llamada de C2	Incluso los ataques sofisticados basados en la IA tienen como objetivo infectar un dispositivo con malware. Ese malware sigue necesitando "llamar" a un servidor C2 para recibir más instrucciones.	Incluso si un dispositivo ya está infectado, el filtrado de DNS puede reconocer y bloquear las consultas enviadas a los servidores C2 para evitar que ejecute su carga malintencionada.
Dominios recién vistos y generados por algoritmos	Los atacantes pueden utilizar la IA para generar dominios únicos y de corta duración como infraestructura para eludir las listas negras estáticas y ejecutar varias etapas de una campaña (p. ej., devoluciones de llamada C2).	El filtrado de DNS clasifica y bloquea las consultas a estos dominios. Los proveedores como Cloudflare, con un volumen y una frecuencia elevados de tráfico DNS, destacan en la detección de estos riesgos.
Tunelización DNS	Los atacantes ocultan el robo de datos codificando los datos confidenciales en consultas DNS de aspecto legítimo. La IA puede facilitar la imitación del tráfico legítimo y evitar la detección en este proceso de codificación, por ejemplo, transmitiendo consultas a intervalos que imiten más la navegación humana por Internet.	El filtrado de DNS utiliza modelos basados en la IA el aprendizaje automático para analizar las propiedades matemáticas, de comportamiento y estructurales de las consultas de DNS a fin de detectar y bloquear los intentos de tunelización.

4 Evita la exposición / exfiltración de datos *(continuación)*

Más allá del filtrado de DNS

Activa los controles HTTP para proteger los flujos de datos cuando los usuarios interactúen con las herramientas de IA

El filtrado de DNS es eficaz por sus sencillas políticas de "permitir o bloquear". Pero para fomentar la adopción de la IA, las organizaciones quieren ir más allá de este enfoque binario y, en su lugar, centrarse en proteger los datos cuando los usuarios interactúan con las herramientas de IA.

Con la inspección HTTP activada, una SWG como Cloudflare puede detectar, bloquear y registrar cualquier intento del usuario de incluir datos confidenciales en una instrucción de IA. Aquí, Cloudflare utiliza las **detecciones clásicas de DPL** para la información de identificación personal (PII), el código fuente, los datos de los clientes, la información financiera, las credenciales y mucho más.

La SWG puede analizar no solo el ***contenido,*** sino también el ***contexto*** de las instrucciones de IA para detener la exposición de datos. En este caso, por ejemplo, Cloudflare busca un propósito inapropiado y malicioso en una instrucción de usuario y crea **medidas de seguridad basadas en el tema y la intención** para evitar cualquier resultado peligroso. Por lo tanto, si una instrucción intenta solicitar información de identificación personal o código malicioso, o intenta eludir las políticas de un modelo de IA, Cloudflare también lo bloqueará y registrará.

La realidad es que la mayoría de las personas compartirán datos en exceso con la IA. De hecho, una encuesta reciente reveló que **el 93 % de los empleados** admite poner información en herramientas de IA sin aprobación.[5] Las detecciones y medidas de seguridad de DLP ayudan a los equipos de seguridad a encontrar el equilibrio entre fomentar la productividad y mitigar los riesgos.

Medidas de protección de instrucciones y seguridad en tiempo real con la SWG de Cloudflare

5 Protege el desarrollo de la IA

Protege a los desarrolladores que crean aplicaciones basadas en la IA

Cada vez más organizaciones no solo adoptan herramientas de IA, sino que también desarrollan internamente sus propias aplicaciones basadas en la IA. La implementación del filtrado de DNS protege a los equipos de desarrolladores responsables de crear estas experiencias de IA en sus flujos de trabajo diarios. El mismo enfoque, que ya ha demostrado su eficacia, puede mitigar nuevos riesgos, entre los que se incluyen los siguientes:

- **Bloqueo del "phishing de modelos" y los intentos de envenenamiento de datos:** las aplicaciones de IA dependen en gran medida de bibliotecas externas, modelos entrenados previamente y datos de centros como Hugging Face e integraciones API. Los atacantes pueden utilizar dominios de suplantación por error tipográfico que alojan servicios de IA falsificados (por ejemplo, el "near-miss" *huggngface.co* en lugar del *huggingface.co* correcto). Los desarrolladores pueden llegar a estos destinos falsos sin darse cuenta, por error al escribir o al hacer clic en un enlace. Una vez allí, se les puede engañar para que inserten credenciales de API, descarguen código malicioso o utilicen modelos y conjuntos de datos envenenados. **El filtrado de DNS interceptaría y bloquearía las consultas a estos dominios peligrosos y, a menudo, nuevos, evitando estos ataques de phishing y a la cadena de suministro.**
- **Prevención de la filtración de los pesos de los modelos:** los pesos de un modelo de IA son sus joyas de la corona, que representan propiedad intelectual de alto valor. Una táctica de exfiltración habitual consiste en utilizar un equipo de desarrollo comprometido para cargar estos archivos en dominios de intercambio de archivos poco conocidos o repositorios privados. **Las políticas de filtrado de DNS adecuadas (p. ej., restringir la resolución de DNS solo a los recursos autorizados) bloquearían la solicitud de la máquina incluso antes de que comenzara cualquier transferencia de datos.**
- **Bloqueo de las inyecciones indirectas de instrucciones:** un desarrollador puede encargar a un agente de IA que analice una página web o contenido que contenga instrucciones ocultas con intenciones maliciosas. Esas instrucciones podrían indicar a la IA que obtenga más datos de un dominio específico o que ejecute una devolución de llamada C2 a un servidor en riesgo. **El filtrado de DNS puede evitar esta inyección indirecta de instrucciones impidiendo que el agente extraiga datos o intente conectarse con el servidor.**

La plataforma para desarrolladores de Cloudflare, en primer plano

Crea experiencias de IA seguras y de primer nivel

La plataforma para desarrolladores de Cloudflare proporciona la infraestructura para escalar tus aplicaciones de IA en cada paso: crea aplicaciones y agentes de IA, almacena datos de entrenamiento, ejecuta inferencia de IA, con seguridad por diseño.

- **Controla y observa las aplicaciones basadas en la IA** y reduce los costes de inferencia y el enrutamiento dinámico del tráfico.
- **Crea servidores de protocolo de contexto de modelo (MCP)** con autenticación y autorización integradas.
- **Evita las interrupciones del servicio** con modelos de recuperación y limitación de velocidad.

Hacia el futuro: cómo garantizar la adopción de la IA con Cloudflare One

Prueba el filtrado de DNS

Como solución independiente, el filtrado de DNS ofrece una forma sencilla y eficaz de abordar los principales desafíos y oportunidades de la IA. Las implementaciones con o sin un cliente de dispositivo y la gestión intuitiva de políticas ayudan a los equipos de seguridad e informáticos a empezar a obtener valor rápidamente en los equipos de trabajo híbridos.

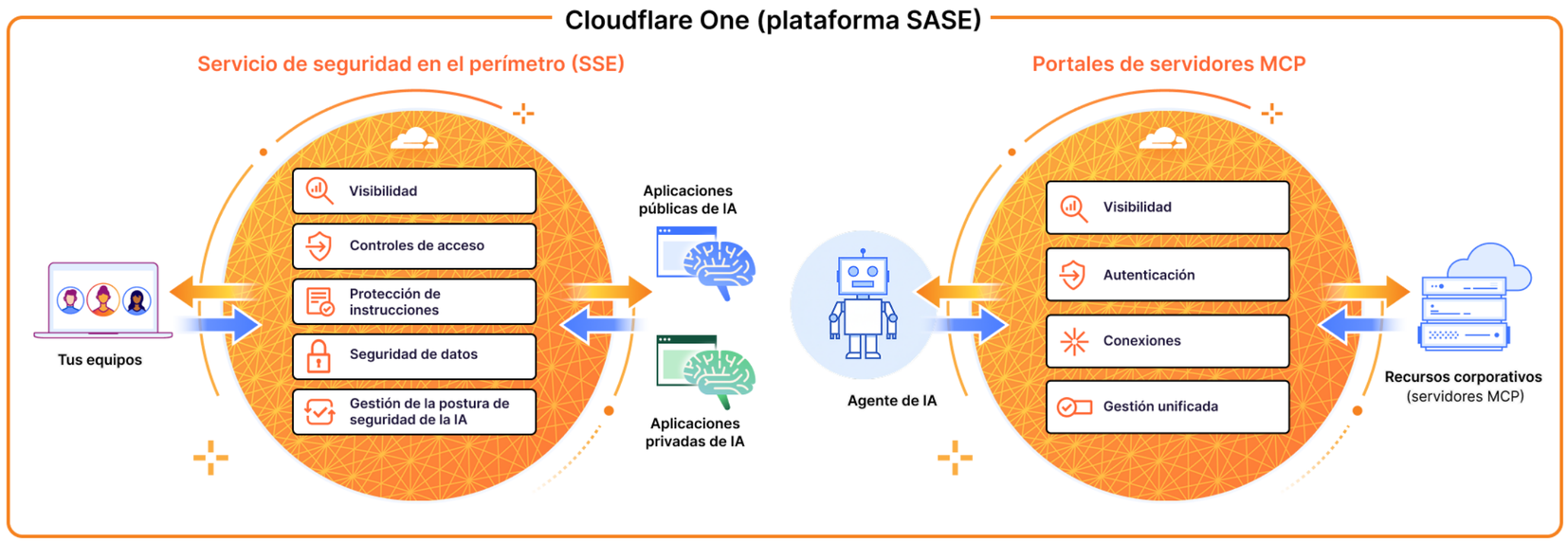
Para muchas organizaciones, la modernización del filtrado de DNS es un primer paso común hacia una implementación completa de SWG o una arquitectura SASE consolidada. Las plataformas como Cloudflare, que agilizan esta progresión, posicionan a tus organizaciones para que se adapten con agilidad y aceleren de forma segura la adopción de la IA.

Más allá del filtrado de DNS

Ampliación de la plataforma SWG y SASE para proteger el uso de la IA generativa y agéntica por parte de los usuarios

Cloudflare One, una plataforma SASE, se sitúa entre tus usuarios y las herramientas de IA, lo que la convierte en un punto de control lógico para proteger el uso de las herramientas de IA. Ya sea que los empleados estén chateando con ChatGPT o que los agentes de IA estén recopilando información a través de los recursos corporativos, Cloudflare ofrece visibilidad y seguridad consistentes en las interacciones entre humanos y la IA y entre máquinas, todo desde un panel de control y un plano de control unificados:

- **Detecta elementos de Shadow AI** y gestiona las políticas para todas las herramientas de IA autorizadas y no autorizadas.
- **Mejora la gobernanza de la IA** con controles de acceso basados en la identidad para el uso de la IA generativa y la comunicación de la IA agéntica.
- **Evita la pérdida de datos** mediante el bloqueo de información confidencial en las instrucciones de los usuarios, la aplicación de controles de temas y la búsqueda de configuraciones incorrectas en las herramientas de IA.



Referencias

1. 2025 IBM, informe Cost of a Data Breach:: https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls
2. Estudio de ManageEngine de 2025: https://www.manageengine.com/survey/shadow-ai-surge-enterprises/
3. Informe CrowdStrike Ransomware: AI Attacks Outpacing Defenses, 2025: https://www.crowdstrike.com/en-us/press-releases/ransomware-report-ai-attacks-outpacing-defenses/
4. “Disrupting the first reported AI-orchestrated cyber espionage campaign,” Anthropic, 13 de noviembre de 2025: https://www.anthropic.com/news/disrupting-AI-espionage
5. Estudios de ManageEngine 2025: https://www.manageengine.com/survey/shadow-ai-surge-enterprises/

+34 518 880 290 | enterprise@cloudflare.com | www.cloudflare.com/es-es/

REV:BDES-8725.28ENE2026