CLOUDFLARE

E-BOOK

Cinq moyens grâce auxquels le filtrage DNS peut vous aider dans votre stratégie de sécurité de l'IA

Sommaire

3 Introduction

4 Identifier l'IA et l'IT clandestines

6 Contrôler l'accès à l'IA

8 Arrêter les cybermenaces assistées par IA

8 Empêcher l'exposition et l'exfiltration des données

10 Protéger le développement de l'IA

11 Aller plus loin : sécuriser l'adoption de l'IA grâce à Cloudflare One

12 Références

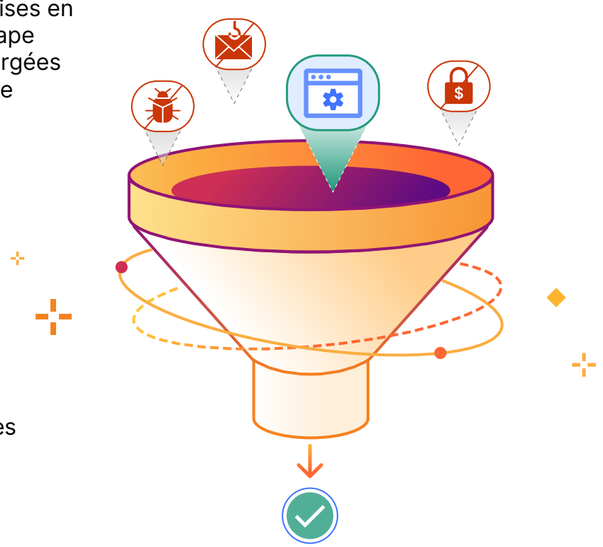
Introduction

Le filtrage DNS assure un délai de rentabilité réduit à vos mesures de sécurité de l'IA

À l'heure où les entreprises s'efforcent d'intégrer l'intelligence artificielle (IA) à leurs flux de travail, l'enthousiasme lié à la volonté de libérer la productivité dissimule souvent des lacunes croissantes en matière de sécurité. L'utilisation non contrôlée d'outils d'IA générative (comme ChatGPT ou Claude) engendre une frontière numérique anarchique au sein de laquelle les données sensibles sont menacées et la conformité peu à peu considérée comme un facteur négligeable. En parallèle, les acteurs malveillants se servent de l'IA comme d'une arme afin de donner un coup de turbo à leurs attaques et d'exploiter cette surface d'attaque en pleine expansion.

Fort heureusement, l'une des technologies de sécurité les mieux établies du secteur, le **filtrage DNS**, peut aider les entreprises à adopter rapidement un moyen plus proactif et plus léger d'atténuer ces risques.

Traditionnellement considéré comme une couche de protection simple et efficace, le filtrage DNS (c'est-à-dire la limitation de l'accès à certains contenus web en fonction du domaine et de l'adresse IP) permet de bloquer les logiciels malveillants qui circulent sur Internet et d'appliquer les politiques d'utilisation acceptable mises en place dans l'entreprise. Il s'agit également d'une étape initiale toujours plus populaire pour les équipes chargées de l'IT et de la sécurité qui débutent le processus de modernisation de leur approche globale en matière de sécurité de l'IA.



Cet e-book présente **cinq moyens courants grâce auxquels le filtrage DNS avec Cloudflare vous aide à adapter votre approche de la sécurité à l'ère de l'IA :**

1. Identifier l'IA clandestine
2. Contrôler l'accès à l'IA
3. Arrêter les cybermenaces assistées par IA
4. Empêcher l'exposition et l'exfiltration des données
5. Protéger le développement de l'IA

À partir de ces bases initiales, les entreprises renforcent souvent leur visibilité et leur contrôle sur d'autres environnements en étendant les fonctionnalités telles que l'inspection HTTP par l'intermédiaire d'une passerelle web sécurisée (SWG, Secure Web Gateway) ou d'une plateforme SASE (Secure Access Service Edge, service d'accès sécurisé en périphérie) plus large. Cet e-book détaille également la manière dont les entreprises peuvent déployer des fonctionnalités SWG et SASE afin de renforcer leur approche de la sécurité de l'IA.

Phase de déploiement avec Cloudflare	Exemple de fonctionnalité
Étape 1 : déployer le filtrage DNS	Analyser l'utilisation de l'IA clandestine et appliquer des mesures de contrôle des accès basées sur le domaine et l'adresse IP.
Étape 2 : approfondir les inspections de la SWG	Bloquer les invites (prompts) des utilisateurs au sein des outils IA en fonction des résultats renvoyés par les mesures de détection des données sensibles et des garde-fous thématiques.
Étape 3 : étendre la plateforme SASE	Appliquer des mesures de contrôle de l'utilisation de l'IA à l'ensemble des communications entre les utilisateurs humains et l'IA, mais aussi des communications entre machines (agents).

1 Identifier l'IA et l'IT clandestines

Filtrez les requêtes DNS pour bénéficier d'une visibilité de base

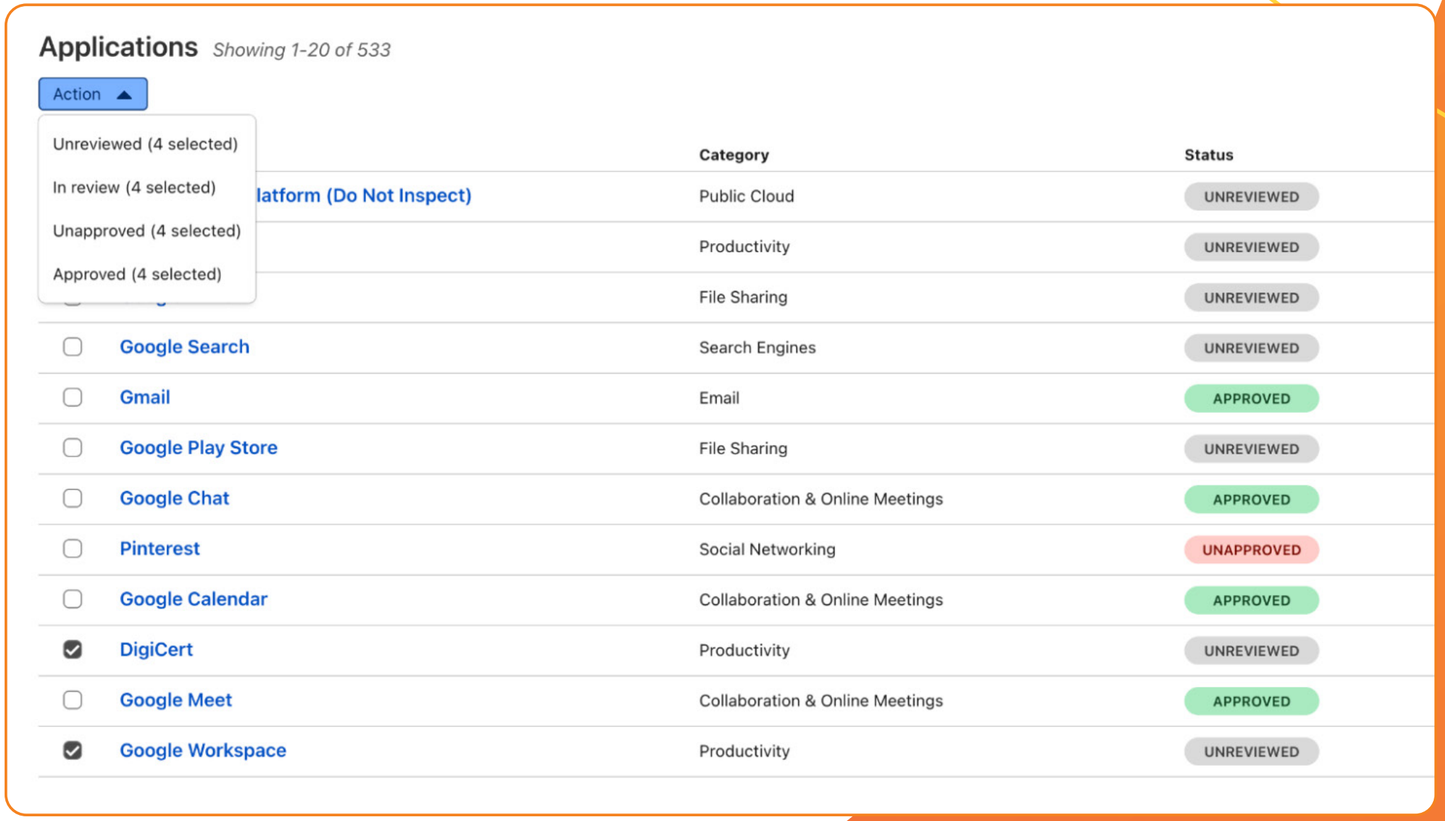
Depuis des années, les entreprises doivent composer avec l'utilisation d'outils SaaS non autorisés ou non approuvés. Or, l'explosion des outils IA et la précipitation autour de leur mise en œuvre sont à l'origine de l'urgence liée à l'IA clandestine qui sévit aujourd'hui :

20 % des entreprises ont subi une violation résultant d'incidents liés à l'IA clandestine (Shadow AI) en 2025.[1]

85 % des responsables informatiques déclarent que leurs collaborateurs adoptent des outils IA avant que le service IT ne puisse les évaluer.[2]

Le filtrage des requêtes DNS vous permet de regagner une visibilité de base sur l'IA clandestine en surveillant chaque requête DNS effectuée par vos utilisateurs. Cette opération vous permet :

- **d'identifier les applications** en fonction de la résolution du domaine (p. ex. chatgpt.com ou clause.ai) ;
- de catégoriser et d'examiner **l'état d'approbation des applications** en fonction du domaine (p. ex. approuvée, non approuvée, non vérifiée ou en cours d'examen). Voir l'exemple à droite ;
- d'évaluer la fiabilité d'une application en fonction de son **score de confiance**. Ce score mesure non seulement les risques généraux posés par les outils SaaS (comme les certifications de conformité et les pratiques de gestion des données), mais également les risques spécifiques à l'IA, notamment ceux qui concernent l'utilisation des données utilisateur pour l'entraînement des modèles ou le fait que le modèle dispose ou non d'une carte système publiée détaillant les tests de biais.



Examinez et identifiez les états des applications dans le tableau de bord

1 Identifier l'IA et l'IT clandestines

Au-delà du filtrage DNS

Affinez vos mesures de contrôle des accès à l'aide de politiques HTTP

Alors que le filtrage DNS propose une base de référence permettant de déterminer **quel trafic est adressé à quelle application**, l'activation de l'inspection HTTP permet de disposer de vues plus granulaires sur **ce que ce trafic effectue au sein de cette application.** Cette visibilité inclut même les journaux des invites et des réponses entre les utilisateurs et les outils d'IA générative.

Les tableaux de bord (comme celui représenté ci-dessous) proposent des analyses agrégées des tendances au fil du temps.

Si vous souhaitez poursuivre l'analyse, cliquez sur n'importe quelle application IA afin de visualiser les utilisateurs ou les groupes spécifiques qui y accèdent, la fréquence à laquelle ils utilisent cette application et leur position géographique, parmi d'autres détails.

Tableau de bord d'analyse des outils d'informatique clandestine

La compréhension des schémas de transfert de données vers et depuis les applications d'IA est une requête courante. **Vous trouverez ci-dessous un exemple d'analyse des données téléchargées/importées par chaque nom d'hôte**. Cette analyse peut ensuite être filtrée par utilisateur, catégorie de contenu et d'autres critères.

Données téléchargées par nom d'hôte

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

Données importées par nom d'hôte

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

2 Contrôler l'accès à l'IA

Définissez des règles d'accès de base en fonction des catégories de domaine

Le filtrage DNS constitue un moyen simple et léger d'empêcher les utilisateurs d'accéder à du contenu Internet malveillant ou indésirable. Pour protéger leurs collaborateurs, les entreprises bloquent généralement l'ensemble des adresses IP et des domaines automatiquement identifiés comme des **risques de sécurité**, tels que les logiciels malveillants, les sites de phishing, les serveurs de type « Command and Control » (C2, prise de contrôle directe), les botnets et les destinations de tunnellisation DNS, par exemple. Elles bloqueront également diverses **catégories de contenu**, comme le contenu pour adultes, les sites de jeux de hasard ou de streaming vidéo, ainsi que certaines **applications particulières spécifiquement définies.** Ce filtrage du contenu sert souvent à appliquer des politiques d'utilisation acceptable pour les collaborateurs ou les clients au sein d'espaces partagés, comme un point de vente, un hôtel, un hôpital ou une école.

Utilisez des catégories de domaine et des sélecteurs d'applications pour contrôler les outils IA auxquels vos utilisateurs peuvent accéder. Vous pouvez, par exemple, associer deux politiques **pour bloquer toutes les applications IA, sauf une que vous auriez approuvée : ChatGPT.**

Étape 1 — **Définir une règle ALLOW pour ChatGPT**

Voir l'exemple de sélecteur

Selector (Required): Application
Operator (Required): in
Valeur: ChatGPT

Étape 2 — **Définir une règle BLOCK pour tous les autres outils IA**

Voir l'exemple de sélecteur

Selector (Required): Catégories de contenu
Operator (Required): in
Valeur: Artificial Intelligence

Les actions de substitution DNS permettent même aux politiques de rediriger le trafic destiné aux domaines à risque vers des ressources internes spécifiques ou des serveurs « gouffres » (sinkholes) en fonction de l'adresse IP. Ainsi, en prenant l'exemple de Cloudflare :

Sélecteur	Opérateur	Valeur	Action	Substitution
Nom d'hôte	IS	www.riskyAI.com	Remplacer	1.2.3.4 (page interne sur la politique en matière d'IA)

2 Contrôler l'accès à l'IA *(suite)*

Au-delà du filtrage DNS

Affinez vos mesures de contrôle des accès à l'aide de politiques HTTP

En activant l'inspection complète du proxy SWG, les entreprises peuvent mettre en place des mesures de contrôle des accès plus précises et plus flexibles par l'intermédiaire de politiques HTTP. Vous trouverez ci-dessous quelques approches populaires en la matière :

- **Appliquer des politiques en matière d'IA clandestine basées sur l'état d'approbation de l'application :** personnalisez les règles selon que les applications sont approuvées/non approuvées/non vérifiées/en cours d'examen. Le blocage de l'ensemble des applications IA non approuvées constitue une option directe, mais vous pouvez également appliquer des actions plus variées, comme celles décrites ci-dessous.
- **Rediriger le trafic vers des URL spécifiques :** vous pouvez, par exemple, envoyer les requêtes utilisateur provenant d'outils d'IA non approuvés vers un outil approuvé ou une page de destination informative.
- **Isoler la session au sein d'un navigateur distant :** routez le trafic destiné aux applications non vérifiées, en cours d'examen ou à d'autres applications spécifiques vers un navigateur isolé au sein duquel l'ensemble du code web s'exécute sur le réseau Cloudflare plutôt que sur un appareil local. Cette mesure d'isolement vous aide à protéger vos données en contrôlant les actions des utilisateurs, notamment en limitant le copier-coller, le téléchargement/l'importation de fichiers, les saisies clavier, etc.
- **Afficher des notifications personnalisées par l'intermédiaire du client sur appareil :** le client sur appareil Cloudflare vous permet, par exemple, d'afficher un message personnalisé lorsque le trafic d'un utilisateur est bloqué. Ce message sert souvent à expliquer la logique à l'origine de la décision de blocage.

Si les exemples ci-dessous constituent autant de politiques d'accès courantes, les politiques HTTP sont nécessaires à la protection plus granulaire des données, notamment les mesures de détection avec prévention des pertes de données (DLP, Data Loss Prevention) examinées dans la section suivante.

3 Bloquer les cybermenaces assistées par IA et 4 Empêcher l'exposition et l'exfiltration des données

Le filtrage DNS demeure efficace contre les menaces émergentes soutenues par l'IA et le vol de données

Les acteurs malveillants font de plus en plus appel à l'IA pour exécuter, automatiser et dimensionner leurs attaques, souvent dans l'objectif classique d'exfiltrer des données sensibles. Ces campagnes peuvent se révéler plus rapides, plus efficaces et plus difficiles à détecter que les autres :

76 % des entreprises reconnaissent avoir du mal à égaler la vitesse et la sophistication des attaques soutenues par IA.[3]

Les chercheurs ont signalé des campagnes dans lesquelles l'IA était à l'origine de **80 à 90 %** des attaques et pour lesquelles seule une intervention humaine minime était requise.[4]

Si les gros titres ont tendance à se concentrer sur les nouvelles techniques de l'IA, comme les deepfakes (hypertrucages) et les logiciels malveillants polymorphes, les acteurs malveillants continuent de s'appuyer sur une infrastructure et des méthodes traditionnelles. Le filtrage DNS constitue ainsi une première ligne de défense efficace contre les menaces issues des deux extrémités de ce spectre.

Le tableau de droite présente les menaces courantes bloquées automatiquement par les services de filtrage DNS et la manière dont elles soutiennent les efforts de vol de données des attaques soutenues par IA. Grâce à sa télémétrie unique, un résolveur DNS récursif et de référence comme celui de Cloudflare, doté d'une visibilité en temps réel sur l'infrastructure Internet mondiale (plus de 5 700 milliards de requêtes DNS par jour), peut notamment soutenir un modèle de recherche de menaces permettant d'identifier ces dernières, souvent à l'aide de l'IA et de l'apprentissage automatique (Machine Learning, ML). La sécurité peut ainsi employer l'IA de manière proactive afin de se protéger contre l'IA.

Menace	Rôle au sein des campagnes soutenues par IA	Utilité du filtrage DNS
Domaines de phishing	L'IA peut générer des appâts hyperpersonnalisés afin d'attirer les cibles vers des sites de phishing qui reposent souvent sur des noms de domaine « d'apparence similaire » (mybank-security.com au lieu de mybank.com, par exemple). Les acteurs malveillants peuvent alors récupérer des identifiants, dérober des cookies de session et pire encore.	La requête échoue avant que la page de phishing ne puisse se charger, même si un collaborateur clique sur un lien de phishing.
Rappels C2	Bon nombre d'attaques cherchent à infecter un appareil avec un logiciel malveillant, même les attaques sophistiquées soutenues par IA. Ce logiciel malveillant doit généralement « rappeler la maison », c'est-à-dire communiquer avec un serveur C2, pour recevoir des instructions supplémentaires.	Le filtrage DNS peut identifier et bloquer les requêtes envoyées aux serveurs C2 afin d'empêcher ces logiciels d'exécuter leur contenu malveillant, même si un appareil est déjà infecté.
Domaines nouvellement observés et générés par algorithme	Les acteurs malveillants peuvent faire appel à l'IA pour générer des domaines uniques et de courte durée en tant qu'infrastructure conçue pour leur permettre de contourner les listes d'exclusion statiques et d'exécuter différentes étapes d'une campagne (comme les rappels C2, par exemple).	Les filtres DNS classent et bloquent les requêtes adressées à ces domaines. Grâce au volume élevé et à la fréquence de leur trafic DNS, certains fournisseurs (comme Cloudflare) excellent dans l'identification de ces risques.
Tunnellisation DNS	Les acteurs malveillants dissimulent le vol de données en encodant des données sensibles au sein de requêtes DNS en apparence légitimes. Grâce à ce processus d'encodage, l'IA peut faciliter l'imitation du trafic légitime et éviter la détection, par exemple, en transmettant des requêtes à des intervalles qui reproduisent plus fidèlement l'activité de navigation Internet d'un être humain.	Les filtres DNS s'appuient sur des modèles soutenus par IA et ML pour analyser les propriétés mathématiques, comportementales et structurelles des requêtes DNS afin de détecter et de bloquer les tentatives de tunnellisation.

4 Empêcher l'exposition et l'exfiltration des données *(suite)*

Au-delà du filtrage DNS

Activez des mesures de contrôle HTTP permettant de protéger vos flux de données lorsque vos utilisateurs interagissent avec des outils IA

Le filtrage DNS est efficace pour ses politiques simples de type « autorisation ou blocage ». Toutefois, afin d'encourager l'adoption de l'IA, les entreprises souhaitent aller au-delà de cette approche binaire et se concentrer à la place sur la protection des données lors de l'interaction des utilisateurs avec des outils soutenus par IA.

Lorsque la fonctionnalité d'inspection HTTP est activée, une SWG telle que celle proposée par Cloudflare peut détecter, bloquer et journaliser toutes les tentatives visant à inclure des données sensibles au sein d'une invite IA effectuées par les utilisateurs. Cloudflare fait ici appel à des **mesures classiques de détection avec DLP** sur les informations d'identification personnelle (PII, Personally Identifiable Information), le code source, les données clients, les informations financières, les informations d'identification et bien plus encore.

La SWG analyse non seulement le ***contenu***, mais également le ***contexte*** des invites IA, afin d'empêcher l'exposition des données. Cloudflare peut ainsi rechercher la présence d'une finalité inappropriée et malveillante dans l'invite d'un utilisateur et génère des **garde-fous basés sur le sujet et l'intention** afin d'empêcher la production de résultats présentant un risque. Ainsi, Cloudflare bloquera et journalisera l'opération si une invite tente de solliciter des informations d'identification personnelle ou du code malveillant (voire de contourner les politiques d'un modèle IA).

La plupart des utilisateurs communiquent trop de données avec l'IA. C'est une réalité. Une récente enquête a ainsi révélé que **93 % des collaborateurs** reconnaissent avoir saisi des informations sans approbation préalable au sein d'outils IA.[5] Les mesures de détection avec DLP et les garde-fous aident les équipes de sécurité à trouver le juste équilibre entre le fait de favoriser la productivité et l'atténuation des risques.

Mesures rapides de protection des invites et garde-fous en temps réel de la solution SWG de Cloudflare

5 Protéger le développement de l'IA

Protégez les développeurs qui conçoivent des applications par IA

De plus en plus d'entreprises adoptent des outils IA, mais développent également des applications assistées par IA en interne. Le déploiement du filtrage DNS protège les flux de travail quotidiens des équipes de développement chargées de la conception de ces expériences IA. Cette approche éprouvée peut atténuer de nouveaux risques, notamment ceux qui figurent ci-dessous.

- **Bloquer le « phishing de modèles » et les tentatives d'empoisonnement de données :** les applications IA dépendent fortement de bibliothèques externes, de modèles pré-entraînés et de données issues de portails du type Hugging Face, ainsi que de l'intégration des API. Les acteurs malveillants peuvent typosquatter des domaines qui hébergent des services IA contrefaits (comme le « presque identique » *huggngface.co* au lieu de la forme correcte *huggingface.co*, par exemple). Les développeurs peuvent se connecter à ces faux sites en orthographiant un lien de manière incorrecte ou en cliquant sur un lien par inadvertance. Ils peuvent alors être amenés à saisir des identifiants API, à télécharger du code malveillant ou à utiliser des modèles et des ensembles de données empoisonnés. **Un filtre DNS permettrait d'intercepter et de bloquer les requêtes adressées à ces domaines dangereux et souvent nouvellement enregistrés afin d'empêcher les tentatives de phishing et les attaques sur la chaîne d'approvisionnement.**
- **Empêcher l'exfiltration des valeurs de pondération des modèles :** les valeurs de pondération d'un modèle IA sont les éléments les plus précieux de ce dernier et constituent une propriété intellectuelle de grande valeur. Une tactique d'exfiltration courante consiste à utiliser une machine de développement compromise pour importer ces fichiers sur des domaines de partage de fichiers obscurs ou des référentiels privés. **Les politiques de filtrage DNS appropriées (restreindre la résolution DNS aux ressources autorisées, par exemple) bloqueront la requête d'une machine avant qu'un transfert de données ne commence.**
- **Blocage des injections indirectes d'invites :** un développeur peut charger un agent IA d'analyser une page web ou du contenu contenant des instructions dissimulées avec une intention malveillante. Ces instructions peuvent inviter l'IA à récupérer davantage de données sur un domaine spécifique ou à procéder à un rappel C2 vers un serveur compromis. **Les filtres DNS peuvent prévenir cette injection indirecte d'invites en empêchant les tentatives d'extraction de données ou « d'appel à la maison » effectuées par l'agent.**

Pleins feux sur la Cloudflare Developer Platform

Développez des expériences IA sécurisées et d'envergure mondiale

La Cloudflare Developer Platform (la plateforme de développement proposée par Cloudflare) vous assure l'infrastructure nécessaire au dimensionnement correct de vos applications IA à chaque stade (développement d'applications et d'agents IA, stockage de données d'entraînement, exécution de l'inférence IA), tout en vous permettant de bénéficier d'une sécurité intrinsèque.

- **Contrôlez et observez vos applications soutenues par IA**, tout en réduisant vos coûts d'inférence et en routant votre trafic de manière dynamique.
- **Concevez des serveurs MCP** (Model Context Protocol, protocole de contexte de modèle) disposant de fonctions d'authentification et d'autorisation intégrées.
- **Prévenez les épisodes d'interruption de service** grâce à des modèles de secours et au contrôle du volume des requêtes.

Aller plus loin : sécuriser l'adoption de l'IA grâce à Cloudflare One

Commencez par mettre un filtrage DNS en place

En tant que solution autonome, le filtrage DNS propose un moyen simple et efficace de s'orienter face aux principales problématiques et opportunités de l'IA. Les déploiements avec ou sans client sur appareil, de même que la gestion intuitive des politiques, aident les équipes chargées de la sécurité et de l'IT à générer rapidement de la valeur pour leurs collaborateurs en travail hybride.

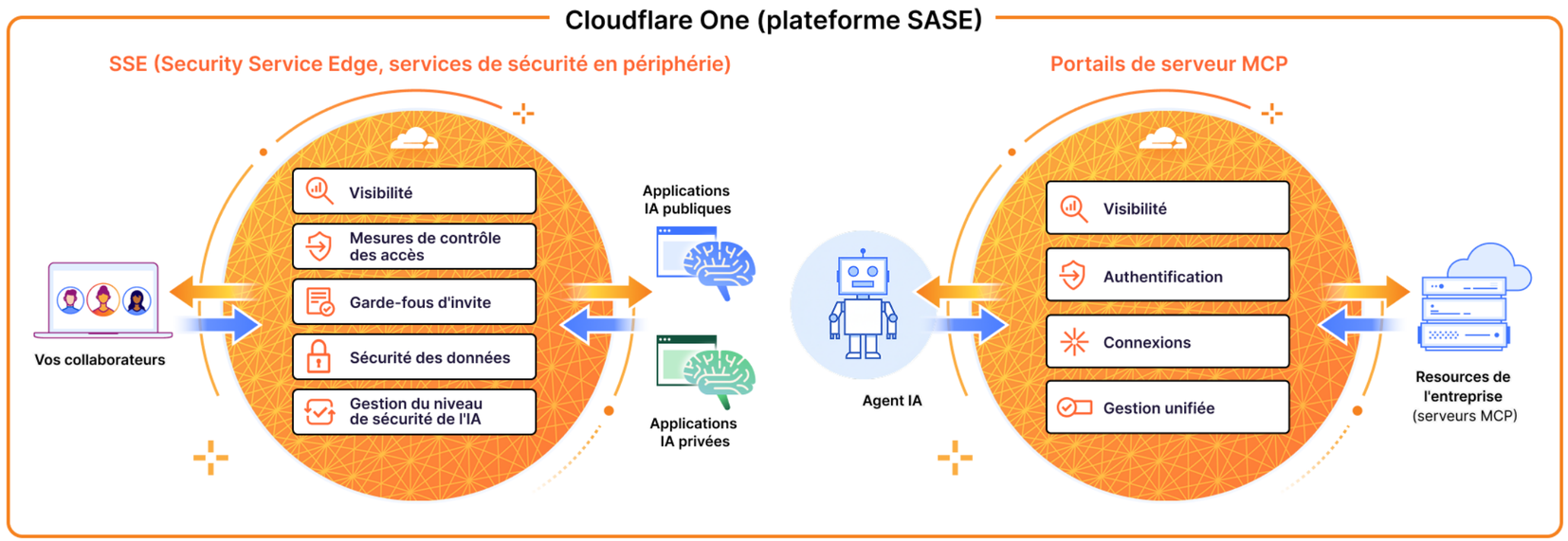
Pour de nombreuses entreprises, la modernisation du filtrage DNS constitue une première étape courante vers le déploiement complet d'un SWG ou une architecture SASE consolidée. Les plateformes qui rationalisent cette progression (comme Cloudflare) permettent à vos entreprises de s'adapter avec agilité et d'accélérer l'adoption de l'IA en toute sécurité.

Au-delà du filtrage DNS

Étendez vos plateformes SWG et SASE afin de sécuriser la manière dont vos collaborateurs utilisent l'IA générative et agentique

Notre plateforme SASE, **Cloudflare One**, se place entre vos collaborateurs et vos outils IA. Elle devient dès lors un point de contrôle logique pour protéger l'utilisation de ces derniers. Que vos collaborateurs dialoguent avec ChatGPT ou que vos agents IA réunissent des informations sur les ressources de l'entreprise, Cloudflare vous assure une visibilité et une sécurité cohérentes sur les interactions entre les utilisateurs humains et l'IA, mais aussi sur les communications entre machines, le tout depuis une interface et un tableau de bord unifiés.

- **Identifiez l'IA clandestine** et gérez les politiques pour tous vos outils IA, autorisés et non autorisés.
- **Renforcez la gouvernance de l'IA** à l'aide de mesures de contrôle des accès basées sur l'identité pour l'utilisation de l'IA générative et la communication agent-IA.
- **Empêchez les pertes de données** en bloquant les informations sensibles contenues dans les invites des utilisateurs, en appliquant des garde-fous thématiques et en recherchant les erreurs de configuration au sein des outils IA.



Références

1. 2025 IBM, Cost of a Data Breach report (rapport 2025 sur le coût d'une violation de données) : 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. Recherches ManageEngine 2025 : https://www.manageengine.com/survey/shadow-ai-surge-enterprises/
3. CrowdStrike Ransomware Report: AI Attacks Outpacing Defenses, 2025 (rapport CrowdStrike 2025 sur les rançongiciels : les attaques soutenues par IA surpassent les mesures de défense) : https://www.crowdstrike.com/en-us/press-releases/ransomware-report-ai-attacks-outpacing-defenses/
4. « Disrupting the first reported AI-orchestrated cyber espionage campaign » (Perturber la première campagne recensée d'espionnage orchestré par l'IA) Anthropic, 13 novembre 2025 : https://www.anthropic.com/news/disrupting-AI-espionage
5. Recherches ManageEngine 2025 : https://www.manageengine.com/survey/shadow-ai-surge-enterprises/

01 73 01 52 44 | enterprise@cloudflare.com | www.cloudflare.com/fr-fr/

RÉV. : BDES-8725.2026JAN28