Éléments importants pour la création d'un environnement d'IA prêt pour la production

Red Hat

Sommaire

Les données : une ressource essentielle pour les entreprises

État du marché de l'IA pour les entreprises

Pour de nombreuses entreprises, l'intelligence artificielle générative (IA générative) est devenue un outil du quotidien.

Les équipes l'utilisent pour résumer des contenus, pour faciliter la création de code et de contenus, ainsi que pour interagir avec les données de façons plus naturelles. À l'échelle de l'entreprise, l'équipe de direction mise sur l'IA générative dans le but d'améliorer les résultats pour les clients, le personnel et l'ensemble de l'exploitation, et pas seulement pour répondre à des questions ponctuelles ou créer des mèmes amusants.

Sur la base des données et applications existantes, l'IA générative peut aider les entreprises à réaliser les actions suivantes :

- Transformer de grands volumes de contenus non structurés en connaissances consultables et réutilisables
- Faciliter le travail des équipes de développement, d'analyse et de rédaction en accélérant la génération et l'amélioration du code, des rapports et des contenus
- Personnaliser l'expérience numérique des clients et du personnel sur tous les canaux
- Automatiser les décisions et workflows ordinaires qui suivent des politiques claires
- Améliorer la productivité des équipes de développement, d'exploitation et métier

D'après de récentes études menées sur ce sujet, cette transition a déjà été amorcée. Le cabinet d'analyse IDC indique que plus de la moitié des entreprises interrogées exécutent déjà plusieurs applications ou services améliorés par l'IA générative en production. Il s'attend à ce que les dépenses sur un an en matière d'IA augmentent d'environ un tiers entre 2025 et 2029, pour atteindre près de 1 300 milliards de dollars d'ici 2029[1]. Pour la plupart des entreprises, l'IA générative fait désormais partie des produits et services de base.

[1] Livre blanc d'IDC, « Agentic AI to Dominate IT Budget Expansion Over Next Five Years, Exceeding 26% of Worldwide IT Spending, and $1.3 Trillion in 2029, According to IDC », 26 août 2025

Parallèlement, les entreprises se préparent à passer à l'étape suivante : l'IA agentique. Au lieu de traiter l'IA générative comme un chatbot ou un assistant unique, l'IA agentique utilise des agents intelligents capables d'appeler des outils, d'interagir avec des applications et de coordonner des tâches composées de plusieurs étapes. Dans la pratique, cette approche peut changer la façon de développer et d'exploiter des logiciels, des fonctions en libre-service pour les clients en passant par l'exploitation informatique et les workflows d'entreprise complexes.

D'après IDC, plus de la moitié des entreprises exécutent déjà des preuves de concept ou des cas d'utilisation précoces de l'IA agentique, et près d'un tiers des applications basées sur l'IA s'appuieront sur cette technologie d'ici fin 2026[2]. Les entreprises la considèrent désormais comme une opportunité stratégique.

Pour bénéficier de cette valeur, les entreprises doivent pouvoir choisir la méthode et l'environnement d'exécution de l'IA.

De nombreuses entreprises prévoient de mettre en place une infrastructure d'IA hybride associant des clouds publics et des environnements sur site dédiés. IDC a constaté que la combinaison hybride d'un cloud public et d'un environnement sur site est devenue la stratégie d'infrastructure numérique la plus courante, et que la plupart des décideurs considèrent que leurs charges de travail d'IA nécessitent un déploiement hybride[3].

Avec une plateforme hybride et ouverte, les entreprises peuvent :

- Garder le contrôle des données et modèles sensibles
- Respecter les exigences en matière de confidentialité et de souveraineté des données
- Faire leur choix parmi toute une gamme d'options matérielles
- Sélectionner des modèles Open Source parmi un large éventail
- Tirer parti de l'évolutivité du cloud en fonction de leurs besoins

Ce livre numérique détaille les principales étapes à suivre pour créer une plateforme d'IA prête pour la production, ainsi que les éléments importants à prendre en compte lors du processus. Il explique aussi l'utilité de la solution unifiée qu'offre Red Hat® AI Enterprise.

[2] Livre blanc d'IDC, « Agentic AI Impact on Digital Infrastructure Strategies », document n° US53418526, octobre 2025 (accès payant)

[3] Livre blanc d'IDC, « AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025 », document n° US53418426, octobre 2025 (accès payant)

Cas d'utilisation de l'IA par secteurs

Santé
- Augmentation de l'efficacité des établissements médicaux
- Hausse de la rapidité et de la précision des diagnostics
- Amélioration des résultats pour les patients

Télécommunications
- Informations sur les comportements des clients
- Amélioration des expériences client
- Optimisation des performances du réseau 5G

Assurance
- Automatisation du traitement des demandes
- Proposition de services d'assurance basés sur l'utilisation
- Aide au calcul des risques

Services financiers
- Personnalisation des services pour les clients
- Amélioration de l'analyse des risques
- Détection des fraudes et du blanchiment d'argent

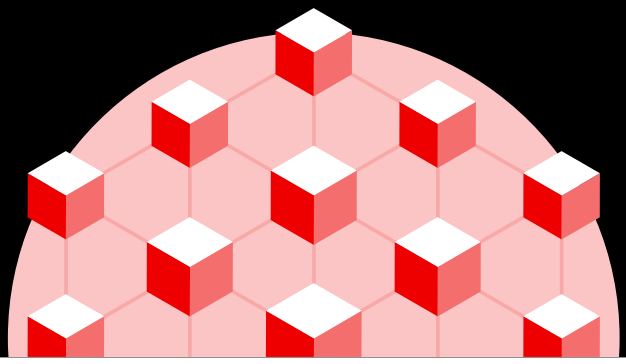
Industrie automobile
- Aide au développement de la conduite autonome
- Prévision des besoins d'entretien
- Amélioration des chaînes d'approvisionnement

Énergie
- Prévision des besoins d'entretien
- Optimisation des interventions et de la sécurité sur le terrain
- Simulation et prévision de réservoir plus rapides

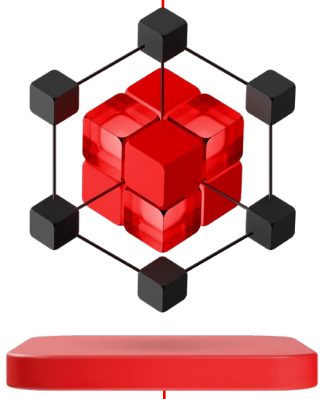
Créer des blocs d'IA pour les entreprises

Ce livre numérique explique comment différents types d'IA fonctionnent ensemble dans une architecture d'entreprise.

- L'**IA générative** utilise de grands modèles de langage (LLM) pour générer du texte, du code et d'autres contenus à partir de données et d'instructions génératives, permettant ainsi aux équipes de travailler plus vite et d'expérimenter des modèles plus facilement.
- L'**IA prédictive** utilise des données historiques et en temps réel afin d'estimer les résultats futurs, comme la demande, les risques ou l'intégrité des équipements, pour que les entreprises puissent agir plus tôt et avec plus de confiance.
- L'**IA agentique** utilise des agents intelligents capables d'appeler des outils, de se connecter à des applications et de coordonner des workflows composés de plusieurs étapes pour atteindre un objectif, plutôt que de répondre à une seule question.
- L'**inférence d'IA** correspond à la phase d'exécution en production, lorsque les modèles appliquent ce qu'ils ont appris à de nouvelles données concrètes pour formuler des prédictions, des recommandations ou des actions. L'inférence peut s'exécuter dans l'ensemble de l'environnement hybride : sur site, dans le cloud ou à la périphérie du réseau.



Création d'une plateforme d'IA prête pour la production



Le développement d'applications basées sur l'IA générative et d'agents intelligents est un processus itératif qui va au-delà de la simple création de modèles d'IA. Le cycle de vie de l'IA se compose des principales étapes suivantes :

1. Définition du cas d'utilisation, détermination des objectifs métier pour le projet d'IA et obtention de l'adhésion des parties prenantes et des responsables
2. Choix de l'environnement dans lequel exécuter les plateformes d'expérimentation et de déploiement de modèles : sur site ou dans le cloud
3. Sélection des modèles d'IA les plus adaptés aux besoins, notamment des modèles ouverts pour éviter toute dépendance
4. Personnalisation des modèles choisis ou alignement sur les données propriétaires à l'aide de la génération augmentée de récupération (RAG)
5. Déploiement des modèles sur un serveur d'inférence
6. Création d'applications ou de charges de travail basées sur l'IA générative
7. Lorsque l'environnement est fonctionnel, utilisation de l'IA agentique pour automatiser le workflow
8. Surveillance et gestion des modèles en toute sécurité et à grande échelle

Avec une architecture d'IA ouverte et adaptable, ce processus peut s'exécuter plus efficacement. Ce type d'architecture nécessite plusieurs technologies et fonctionnalités clés :

- **Accès aux modèles de type Open Weight et frontier**, qui offrent un point de départ aux entreprises
- **Outils GenAIOps et DevOps**, qui permettent aux équipes d'ingénierie de l'IA, de science des données, d'ingénierie de l'apprentissage automatique (AA) et de développement d'applications de créer, déployer et gérer des modèles d'IA, des agents intelligents et des applications basées sur l'IA
- **Accès aux outils de réglage des modèles, comme le réglage fin et les fonctionnalités de RAG**, pour personnaliser les modèles avec des données d'entreprise privées et s'adapter aux cas d'utilisation spécifiques d'un domaine
- **Environnements d'exécution de l'inférence**, qui permettent d'obtenir un niveau optimal de performances, de débit et de latence
- **Composants fondamentaux pour les agents intelligents** afin de gérer, gouverner et sécuriser leur mise en œuvre en production
- **Accélérateurs de calcul, de stockage et du réseau** pour écourter la préparation des données, la personnalisation des modèles et les tâches d'inférence
- **Points de terminaison d'infrastructure**, qui fournissent des ressources pour les environnements de cloud privé, public et hybride, sur site, virtuels et d'edge computing pour toutes les étapes de l'exploitation de l'IA

Ce livre numérique présente les éléments essentiels à prendre en compte pour la création d'une architecture d'IA efficace.

L'inférence correspond à la phase d'exécution en production de l'IA. Un modèle sera inutile tant qu'il n'aura pas d'API et qu'il ne distribuera pas de contenu. Ce contenu est distribué lors de l'inférence.

Chris Wright
Directeur technique de Red Hat[4]

[4] Ron Miller, « Red Hat's CTO sees AI as next step for company's open approach », Fastforward, 11 novembre 2025

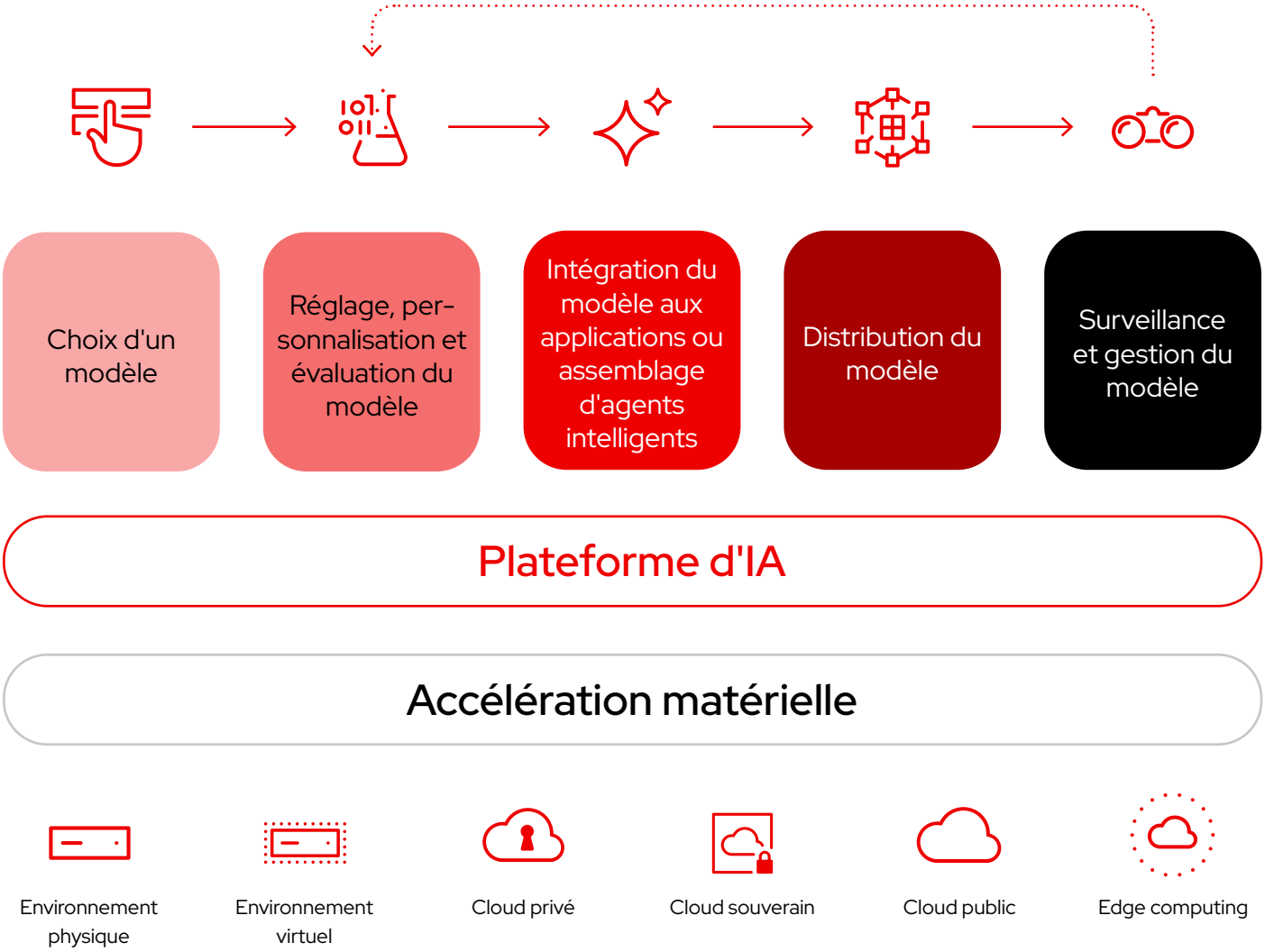


***Figure 1** : les composants d'une architecture d'IA*

Défis liés au déploiement de l'IA

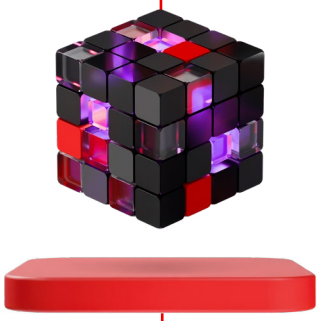
Les entreprises doivent choisir, assembler et fournir des solutions d'IA qui offrent un avantage concurrentiel. Plusieurs défis freinent la mise en œuvre et à l'échelle des déploiements d'IA :

- **Coût du modèle** : l'exécution de grands modèles et de l'inférence à grande échelle peut avoir un coût élevé. Les entreprises doivent optimiser les modèles et l'inférence pour limiter les coûts de calcul tout en fournissant des applications précises et réactives.
- **Complexité de l'alignement** : l'entraînement et le réglage de modèles, ainsi que la création de pipelines de RAG, sont des tâches complexes qui sollicitent énormément les processeurs graphiques (GPU). Les entreprises peuvent simplifier la personnalisation de leurs données et faire appel à des spécialistes et des développeurs pour passer plus rapidement des expérimentations à la production.
- **Contrôle et cohérence** : les services d'IA intégrés limitent le contrôle sur le matériel, les données et la gouvernance. Les entreprises peuvent opter pour une approche hybride, permettant de sélectionner des modèles et une infrastructure tout en conservant la propriété des données, du cycle de vie et de la mise à l'échelle des déploiements.

Pour relever ces défis, une plateforme d'IA ouverte et hybride doit être mise en place. Celle-ci fournira des outils cohérents pour l'optimisation, la personnalisation et la gouvernance des modèles dans tous les environnements.

Éléments à prendre en compte pour la plateforme d'IA

Conteneurs et orchestration des conteneurs



Conteneurs

Un conteneur est une unité de base d'un logiciel qui contient des applications avec toutes leurs dépendances. Les conteneurs simplifient les processus de création d'applications et permettent leur déploiement dans différents environnements, sans modifications.

Importance pour l'IA

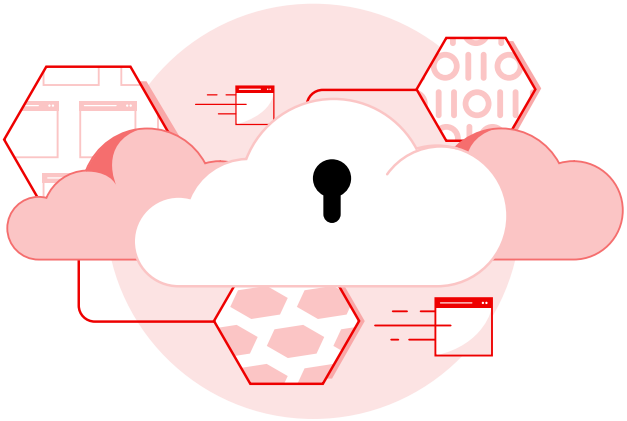
D'un côté, les équipes d'ingénierie de l'IA et de développement d'applications ont besoin d'accéder aux outils et ressources de leur choix pour optimiser leur productivité. De l'autre, les équipes d'exploitation doivent s'assurer que les ressources sont à jour, conformes et utilisées de manière sécurisée.

Parce qu'ils regroupent les serveurs de modèles, les dépendances et les configurations dans une unité reproductible facilitant la gestion des mises en production, à l'échelle et à jour, les conteneurs constituent souvent la meilleure option pour déployer des LLM et des applications basées sur l'IA générative.

Les conteneurs permettent de déployer une large sélection d'outils d'IA dans des environnements hybrides, et ce de manière cohérente. Les équipes peuvent modifier de manière itérative et partager des images de conteneurs grâce au système de contrôle de version qui suit les changements entre les versions pour plus de transparence. En parallèle, les fonctionnalités d'isolation des processus et de contrôle des ressources améliorent la protection contre les menaces.

Meilleures pratiques et recommandations

Cherchez une plateforme de conteneurs flexible et hautement disponible qui offre des fonctions de sécurité intégrées et qui rationalise le déploiement, la gestion et les déplacements des conteneurs dans l'ensemble de l'environnement. Sélectionnez une plateforme Open Source compatible avec un large éventail de technologies pour profiter de plus de choix et de flexibilité.



Orchestration des conteneurs

L'orchestration des conteneurs est le processus qui consiste à gérer la création, le déploiement et le cycle de vie des conteneurs dans l'ensemble de l'environnement.

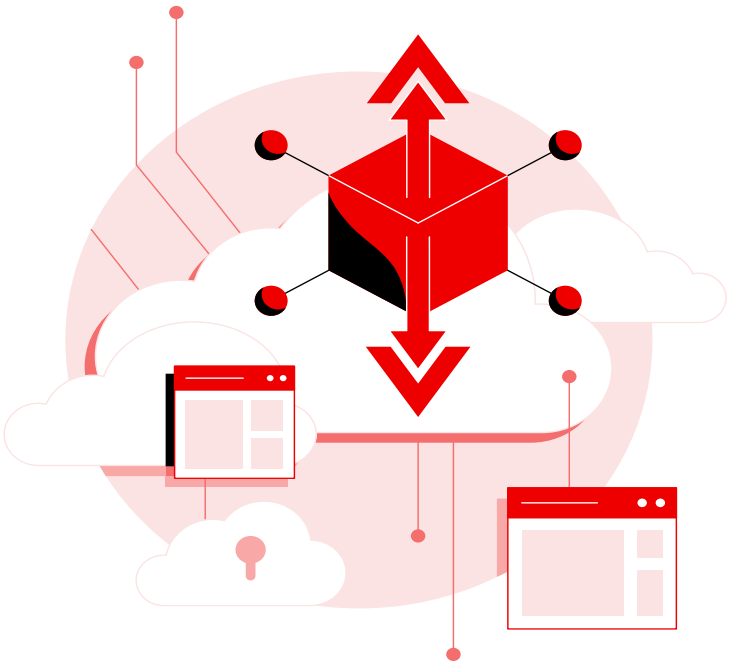
Importance pour l'IA

Après avoir adopté des conteneurs, il faut un moyen efficace pour les déployer, les gérer et les faire évoluer. Les moteurs d'orchestration des conteneurs permettent d'administrer le cycle de vie des conteneurs de manière cohérente. Ce type d'outil centralise l'accès aux ressources de calcul, de stockage et de réseau dans les environnements sur site, cloud et d'edge computing. Ces outils fournissent également des fonctions unifiées d'ordonnancement des charges de travail, de contrôle multi-client et d'application de quotas.

Meilleures pratiques et recommandations

Choisissez un environnement d'orchestration des conteneurs basé sur Kubernetes qui s'appuie sur une technologie Open Source de pointe et permet d'éviter toute dépendance vis-à-vis d'un cloud propriétaire. Cherchez une plateforme qui offre un contrôle multi-client renforcé, un accès basé sur les rôles et des fonctions de gestion des politiques afin de gérer les charges de travail d'IA de manière cohérente. Privilégiez les solutions associées à un vaste écosystème d'opérateurs et d'intégrations afin de standardiser le déploiement, la mise à l'échelle et la gestion des services d'IA dans l'ensemble des environnements hybrides.

D'ici 2027, plus de 75 % de tous les déploiements d'IA devraient utiliser des conteneurs comme environnement de calcul sous-jacent, contre moins de 50 % en 2024.[5]



[5] Gartner, « Magic Quadrant for Container Management », 10 septembre 2024

Éléments à prendre en compte pour la plateforme d'IA

Gestion des applications et pratiques GenAIOps

Gestion du cycle de vie des charges de travail d'IA

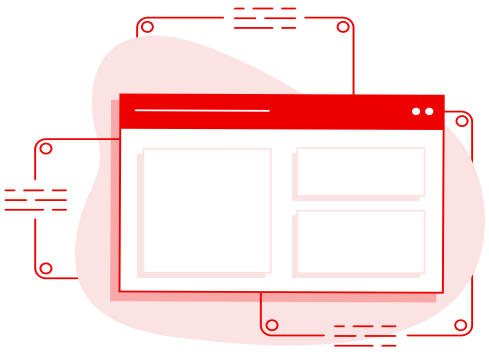
La gestion du cycle de vie des charges de travail d'IA couvre le déploiement, la mise à l'échelle et l'administration des outils et services qui permettent l'exécution des cas d'utilisation de l'IA.

Importance pour l'IA

Par nature, les environnements d'IA sont complexes. Les composants de gestion du cycle de vie des charges de travail d'IA (comme les notebooks, les workbenches, les pipelines et les points de terminaison de distribution des modèles) doivent être conteneurisés pour simplifier le contrôle et la gestion. Les équipes d'exploitation peuvent automatiser des tâches courantes du cycle de vie telles que la configuration, le provisionnement et les mises à jour pour améliorer la précision et limiter les tâches manuelles. Les équipes de science des données, d'ingénierie de l'IA et de développement d'applications peuvent demander un environnement d'IA préapprouvé dans un catalogue sans ouvrir de tickets auprès du service informatique. L'automatisation permet également de consacrer plus de temps aux activités stratégiques à plus forte valeur ajoutée, plutôt qu'à des tâches répétitives.

Meilleures pratiques et recommandations

Pour une gestion efficace du cycle de vie des charges de travail d'IA, commencez avec une sélection d'images de workbench et de notebook incluant des bibliothèques d'IA et d'AA couramment utilisées, qui permettra aux équipes d'utiliser une base de référence sécurisée et prise en charge plutôt que des environnements ponctuels. Les entreprises doivent fournir des environnements de notebooks basés sur un navigateur avec intégration de Git, pour permettre aux équipes de collaborer sur des expérimentations et de suivre les modifications apportées au code et aux modèles au fil du temps.



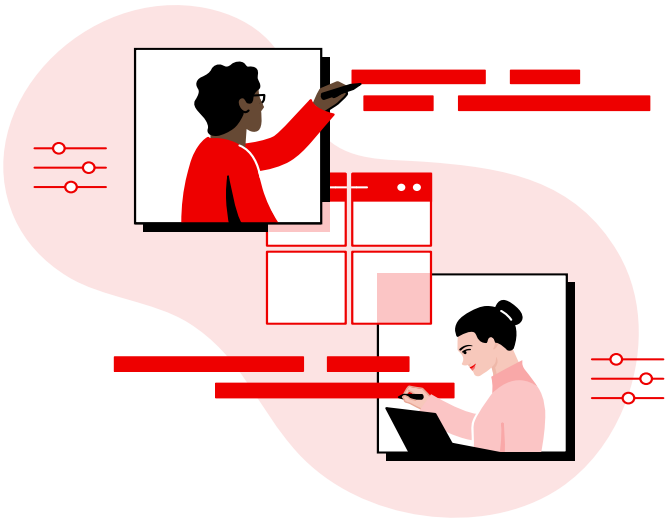
Pratiques GenAIOps et MLOps

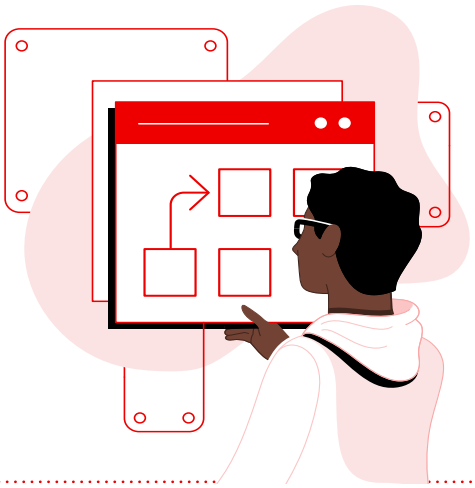
Les pratiques GenAIOps et MLOps rassemblent les outils, plateformes et processus nécessaires pour mettre en œuvre l'IA à grande échelle.

Importance pour l'IA

Les entreprises doivent développer et déployer rapidement et efficacement des modèles d'IA ainsi que les applications qui les utilisent. La collaboration entre les équipes joue ici un rôle essentiel.

Tout comme le DevOps, les approches GenAIOps et MLOps favorisent cette collaboration entre les équipes d'ingénierie de l'IA, de développement d'applications et d'exploitation informatique, dans le but d'accélérer la création, l'entraînement, le déploiement et la gestion des modèles d'IA générative, des agents intelligents et des applications basées sur l'IA. L'automatisation, souvent sous la forme de pipelines de CI/CD (intégration et distribution continues), permet d'effectuer des changements rapides, progressifs et itératifs qui accélèrent les cycles de développement des applications et des modèles.





Pratiques GenAIOps et MLOps

En plus des technologies, les pratiques GenAIOps et MLOps englobent les équipes et des processus. Les pratiques GenAIOps et MLOps peuvent être appliquées à l'ensemble du cycle de vie de l'IA. L'utilisation de l'automatisation sur les plateformes et dans les outils, ainsi que de technologies Open Source comme Kubeflow, permet de créer des pipelines et workflows de CI/CD.

Éléments à prendre en compte pour la plateforme d'IA

Plateforme de cloud hybride

Une plateforme de cloud hybride fournit une base pour le développement, le déploiement et la gestion de l'IA dans des environnements sur site, cloud et d'edge computing. Elle permet également de concevoir une IA souveraine et une IA privée dès le début. Les entreprises peuvent ainsi choisir les charges de travail à exécuter dans des clouds publics et celles qui resteront sur site ou dans un cloud privé qu'elles contrôlent.

Importance pour l'IA

Les modèles, agents, logiciels et applications d'IA nécessitent une infrastructure évolutive pour leur développement et leur déploiement. Avec une plateforme de cloud hybride cohérente, il est possible de développer, régler, tester, déployer et gérer des modèles et applications d'IA de la même manière dans toutes les parties de l'infrastructure, pour plus de flexibilité.

Ce type de plateforme soutient également les stratégies d'IA souveraine et d'IA privée. Les entreprises peuvent conserver leurs données et modèles sensibles dans des régions spécifiques ou même des environnements déconnectés, afin de répondre aux exigences en matière de résidence, de confidentialité et de conformité des données, tout en gardant la possibilité de se connecter aux services de cloud public en cas de nécessité. Le déploiement de fonctionnalités en libre-service peut accélérer la distribution des ressources tout en préservant leur contrôle.

Enfin, une plateforme cohérente fournit une base pour intégrer des solutions technologiques de fournisseurs tiers et de communautés Open Source, ainsi que tous les outils personnalisés et nécessaires.

Meilleures pratiques et recommandations

Choisissez une plateforme centrée sur la sécurité qui prend en charge l'accélération matérielle, qui s'appuie sur un vaste écosystème d'outils de développement d'applications et d'IA et qui intègre des fonctionnalités de gestion de l'exploitation et GenAIOps.

Cherchez une solution qui offre un contrôle strict des politiques en matière de localisation des données, de placement des modèles et d'accès afin d'exécuter des charges de travail d'IA souveraine et privée sur site ou dans des clouds privés, tout en gardant la possibilité de se connecter à des clouds publics en cas de nécessité.

Le choix d'une plateforme Open Source peut apporter davantage de possibilités d'intégration et de flexibilité, favorisant l'innovation rapide au travers du développement communautaire. Cette approche peut aussi donner à accès des fonctionnalités en libre-service permettant d'accélérer la distribution des ressources tout en préservant le contrôle de l'environnement informatique.

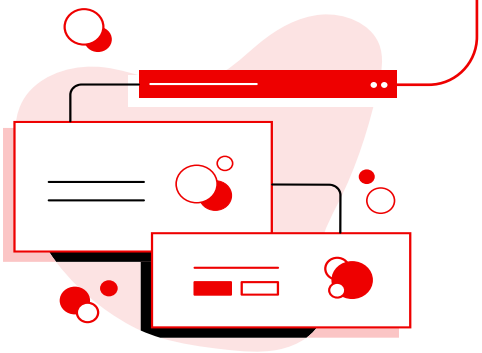
La combinaison hybride d'un cloud public et d'un environnement sur site dédié est la stratégie d'infrastructure numérique la plus courante[3].

[3] Livre blanc d'IDC, « AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025 », document n° US53418426, octobre 2025 (accès payant)

Éléments à prendre en compte pour la plateforme d'IA

Personnalisation et alignement des modèles

Les applications modernes basées sur l'IA nécessitent des modèles qui reflètent les données, workflows et contraintes métier spécifiques de l'entreprise. L'alignement d'un modèle frontier ou ouvert sur les informations propriétaires permet aux entreprises de passer de réponses générales à des résultats précis et adaptés à leur domaine.



Importance pour l'IA

L'IA générative et l'IA agentique dépendent de modèles qui comprennent la terminologie, les données et la situation réelle de l'entreprise.

L'alignement permet de préserver la précision et la pertinence des résultats en basant les modèles sur les données privées. Cette technique améliore l'efficacité en réduisant les coûts d'inférence et en évitant un surdimensionnement inutile. Elle renforce la gouvernance et le contrôle en facilitant la mise en œuvre d'une logique métier, de règles de sécurité et d'exigences de conformité directement dans le comportement du modèle. Elle favorise aussi l'évolutivité en proposant des processus cohérents pour mettre à jour, réentraîner et versionner les modèles selon l'évolution des données.

La personnalisation soutient également les stratégies d'IA souveraine et d'IA privée, en permettant aux entreprises d'entraîner et de distribuer des modèles intégralement dans des environnements contrôlés afin de respecter les exigences réglementaires, de confidentialité et de résidence des données.

Meilleures pratiques et recommandations

Adoptez des workflows modulaires qui commencent par la RAG, le réglage fin, l'ingénierie de prompt et les couches de politiques en fonction de vos besoins plutôt que de vous appuyer sur une seule méthode. Utilisez des modèles ouverts pour éviter toute dépendance et pour garder la possibilité d'optimiser les modèles à l'aide du réglage fin et de la quantification, ainsi que de les évaluer en toute transparence. Impliquez des spécialistes pour vous assurer que les modèles reflètent le contexte métier réel et la précision des données. Optimisez de manière précoce votre modèle pour l'inférence avec des techniques telles que la quantification, la distillation et des environnements d'exécution efficaces pour contrôler les coûts et la latence. Préservez aussi la reproductibilité et l'efficacité de la gouvernance avec des ensembles de données de version, des exécutions d'entraînement, des pondérations de modèles et des indicateurs de mesure d'évaluation.

Éléments à prendre en compte pour la plateforme d'IA

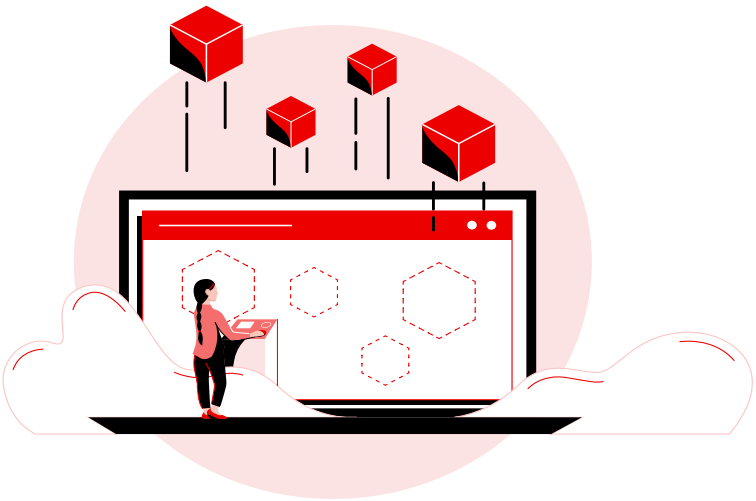
Inférence d'IA à grande échelle

L'exécution de l'IA en production nécessite des opérations d'inférence rapides, efficaces et fiables. L'inférence est la phase qui suit l'entraînement ou l'alignement des modèles. À cette étape, les modèles traitent de nouvelles données, formulent des prédictions, génèrent des contenus ou déclenchent des actions dans une application ou un workflow.

Avec l'adoption de l'IA générative et de l'IA agentique, l'inférence devient un important facteur de coût et de performances, en particulier lorsque les applications passent d'interactions basées sur une seule requête à des tâches continues en plusieurs étapes et exécutées par des agents intelligents.

Importance pour l'IA

Les opérations d'inférence influent directement sur l'expérience utilisateur, les performances des applications et les coûts d'exploitation. Les charges de travail d'IA générative et d'IA agentique nécessitent souvent des réponses rapides, des requêtes parallèles et un débit cohérent dans de nombreux environnements, des datacenters au cloud public et aux sites d'edge computing.



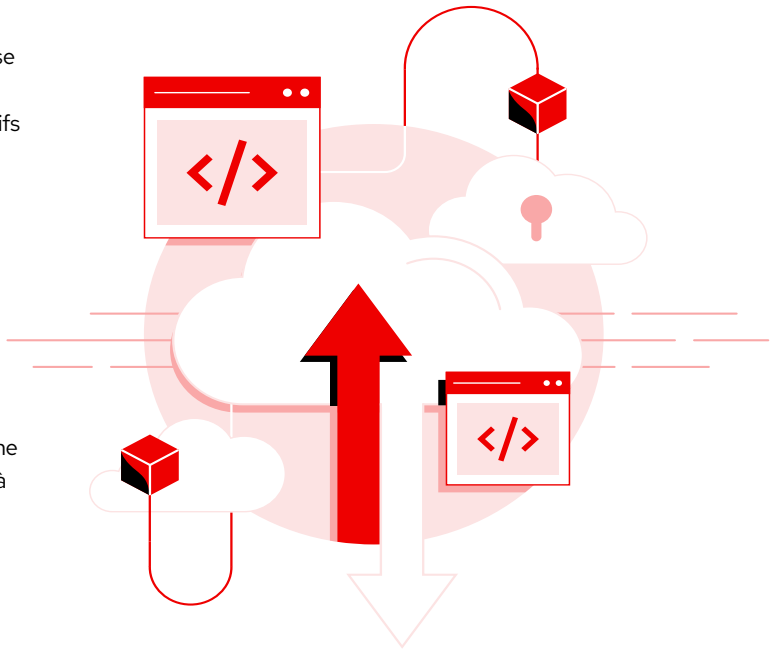
Avec des environnements efficaces d'exécution de l'inférence, il est possible de réduire les coûts liés aux GPU et aux processeurs, d'améliorer la latence pour les tâches interactives, ainsi que de répondre aux besoins en constante évolution des agents intelligents qui appellent des outils, utilisent des interfaces de programmation d'application (API) et coordonnent des workflows composés de plusieurs étapes. L'optimisation de l'inférence soutient aussi les stratégies d'IA souveraine et d'IA privée, en permettant aux entreprises d'exécuter des opérations d'inférence à proximité des données sensibles, sur site ou dans des clouds privés, tout en conservant des performances prévisibles.

Meilleures pratiques et recommandations

Choisissez des environnements d'exécution de l'inférence adaptés à votre type de modèle et à votre environnement de déploiement, que ce soit pour des LLM, des modèles multimodaux, des modèles prédictifs ou des charges de travail agentiques. Privilégiez les environnements d'exécution et l'infrastructure qui prennent en charge la mise à l'échelle dynamique, à la fois horizontale et verticale, afin de répondre aux demandes imprévisibles des LLM interactifs et des opérations d'inférence basées sur des agents.

Utilisez des techniques telles que la quantification, la distillation et l'optimisation des modèles afin de réduire les coûts et d'améliorer la latence. Faites appel à des fournisseurs expérimentés si vous ne disposez pas des compétences nécessaires. Associez ces techniques d'optimisation à des technologies couramment utilisées, comme vLLM pour l'inférence de LLM à haut débit, ainsi qu'à de nouveaux frameworks d'inférence distribuée, comme llm-d, qui désagrège le processus d'inférence pour mettre à l'échelle chaque phase de manière indépendante.

Déployez l'inférence dans des conteneurs pour regrouper les dépendances et permettre une mise à l'échelle cohérente dans les environnements hybrides. Placez des points de terminaison d'inférence où vos données et applications sont stockées pour limiter leur déplacement et garder le contrôle, en particulier dans les cas d'utilisation de l'IA souveraine et privée. Enfin, surveillez les performances des modèles au fil du temps et mettez à jour les versions à mesure que la distribution des données évolue afin de préserver la précision et la fiabilité à grande échelle.



90 %

des décideurs estiment que l'IA va jouer un rôle important dans le budget de leur infrastructure numérique et dans leurs choix technologiques en 2026[3].

[3] Livre blanc d'IDC, « AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025 », document n° US53418426, octobre 2025 (accès payant)

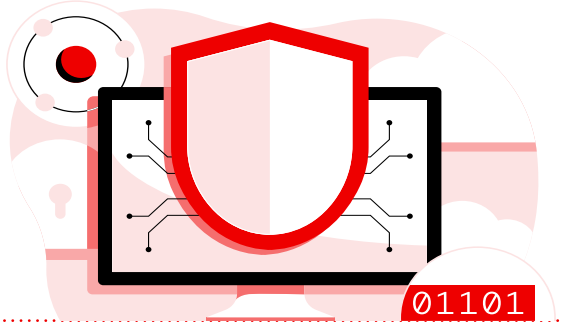
Éléments à prendre en compte pour la plateforme d'IA

Sécurité de l'IA

Les systèmes d'IA doivent se comporter de manière fiable, prévisible et conforme aux politiques de l'entreprise. À mesure que les entreprises passent de l'expérimentation à la production, la sécurité de l'IA devient essentielle, en particulier lorsqu'elles déploient l'IA générative, l'IA agentique et des workflows autonomes capables de réaliser des actions, plutôt que de simplement formuler des suggestions.

Importance pour l'IA

La sécurité de l'IA consiste à maintenir les modèles et agents d'IA dans les limites définies afin de respecter les exigences métier, juridiques et éthiques. En cas de résultats imprécis, d'une dérive de modèle, d'un traitement non sécurisé des données ou d'actions non prévues, de réels risques peuvent affecter l'exploitation de l'IA. L'IA générative et les systèmes agentiques posent aussi de nouvelles questions de sécurité, comme les hallucinations, l'exécution d'outils non approuvés, l'augmentation des privilèges et l'incohérence du raisonnement lors de tâches composées de plusieurs étapes. Grâce à des pratiques de sécurité efficaces, les entreprises peuvent préserver la confiance, protéger les données sensibles et prévenir toute action malveillante ou irréversible. Dans les secteurs réglementés, les contrôles de sécurité sont essentiels pour assurer la conformité et préparer les audits des environnements hybrides et sur site.



Meilleures pratiques et recommandations

Adoptez une approche de sécurité multicouche qui inclut des garde-fous basés sur des politiques, des filtres de contenus et des contrôles d'exécution des outils pour les workflows agentiques. Validez et testez régulièrement les modèles pour détecter les dérives ou la dégradation de la précision. Exécutez les charges de travail sensibles dans des environnements privés ou sur site pour garder le contrôle sur l'exposition des données et le comportement des modèles, conformément aux stratégies en matière d'IA souveraine et d'IA privée. Utilisez des frameworks d'évaluation des modèles pour surveiller les biais, la robustesse et la fiabilité. Cherchez à améliorer vos modèles et vos données à l'aide d'outils qui les stockent dans des registres OCI standards et qui fournissent des chaînes d'approvisionnement sécurisées. Les technologies largement adoptées, comme vLLM pour l'inférence des LLM, et les nouvelles technologies distribuées, comme llm-d, peuvent contribuer à réduire les coûts et à mettre à l'échelle le déploiement de votre projet d'IA. Enfin, versionnez et documentez vos modèles, ensembles de données et politiques afin d'assurer le suivi des décisions et la cohérence de la gouvernance tout au long du cycle de vie de l'IA.

Assemblage d'une base flexible et ouverte pour l'IA

Red Hat AI Enterprise est une plateforme d'IA intégrée pour le développement et le déploiement de modèles, agents et applications d'IA efficaces et rentables dans des environnements de cloud hybride. Elle fait partie de l'offre Red Hat AI.

Cette solution unifie les cycles de vie des modèles et applications d'IA pour augmenter l'efficacité opérationnelle, accélérer la distribution et atténuer les risques en fournissant un environnement de développement prêt à l'emploi avec des fonctionnalités adaptées aux entreprises.

Reposant sur Red Hat OpenShift, cette pile d'IA complète, testée et prise en charge améliore l'interopérabilité et garantit la continuité des activités. Elle inclut des fonctionnalités essentielles comme le réglage de modèles, l'inférence hautes performances et la gestion des workflows d'IA agentique. Elle offre aussi aux équipes la possibilité de choisir les modèles, le matériel et l'environnement de déploiement qui leur conviennent, tout en respectant les exigences en matière d'emplacement des données. La solution Red Hat AI Enterprise est prise en charge dans tous les environnements hybrides, ce qui permet aux équipes de planifier la capacité, les GPU et les futurs projets d'IA en toute confiance.

La solution Red Hat AI Enterprise inclut les technologies issues du projet Open Source llm-d, lancé par Red Hat avec des partenaires comme IBM, NVIDIA, Google et AMD. Le framework llm-d améliore la rentabilité en séparant les phases de préremplissage et de décodage de l'inférence, afin que chacune puisse évoluer différemment. Son module d'équilibrage de charge sensible à l'inférence achemine les requêtes en fonction des files d'attente de jetons textuels, ce qui permet de réduire les délais de réponse et, dans certains cas, de diriger les charges de travail de préremplissage vers les processeurs.

Réduction du délai de rentabilisation
Déployez une pile d'IA optimisée pour les entreprises sur l'infrastructure de votre choix, avec des outils préconfigurés, des déploiements automatisés et des fonctions intégrées d'observabilité. Les équipes de développement et d'ingénierie de l'IA peuvent ainsi se concentrer sur la création et la distribution d'applications cloud-native agentiques et basées sur l'IA.

Amélioration de l'efficacité opérationnelle
Rationalisez et automatisez les workflows, de la validation du code à l'établissement de workflows de pipelines d'IA, en passant par le déploiement de modèles. Les équipes d'exploitation informatique peuvent ainsi fournir des performances stables et exploiter davantage l'infrastructure existante grâce à une allocation intelligente des ressources et à la gestion intégrée du cycle de vie.

Atténuation des risques
Réduisez les risques liés à l'adoption de l'IA pour les entreprises grâce à une pile d'IA intégrée, testée et entièrement prise en charge, qui améliore l'interopérabilité entre tous les modèles, tous les équipements et tous les environnements de cloud hybride. Tirez parti de cette base pour respecter les exigences réglementaires et de résidence des données afin de mettre à l'échelle l'IA en toute confiance.

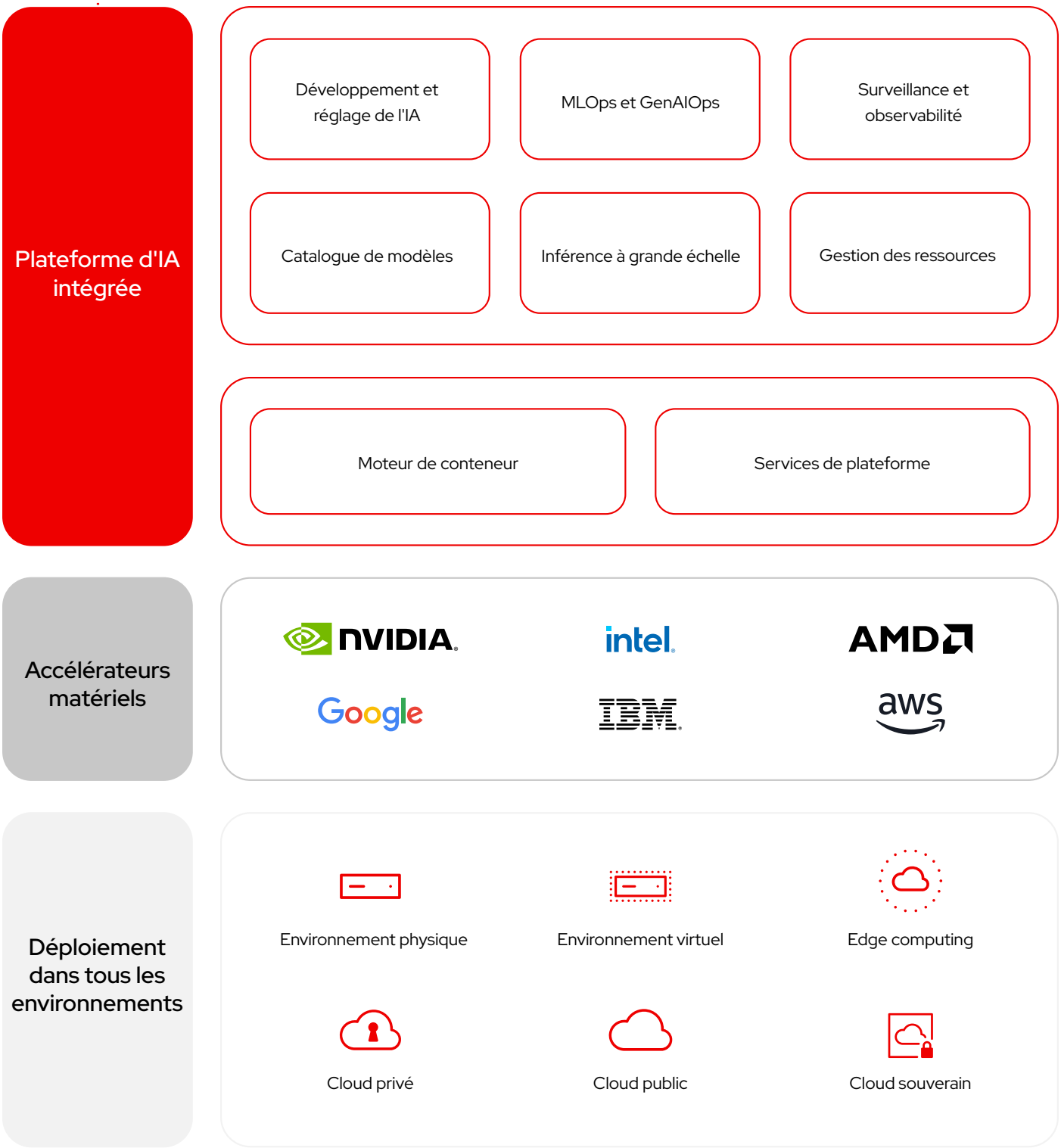
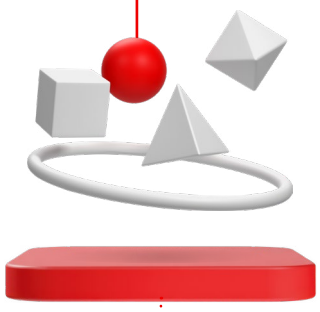


***Figure 2** : les composants d'une plateforme d'IA intégrée*

Plus de choix et de flexibilité avec un écosystème de partenaires certifiés pour l'IA/AA



Avec l'évolution rapide des outils et technologies d'IA, il devient difficile pour les entreprises de suivre le rythme des avancées tout en préservant la stabilité et la fiabilité au sein de leur environnement informatique.

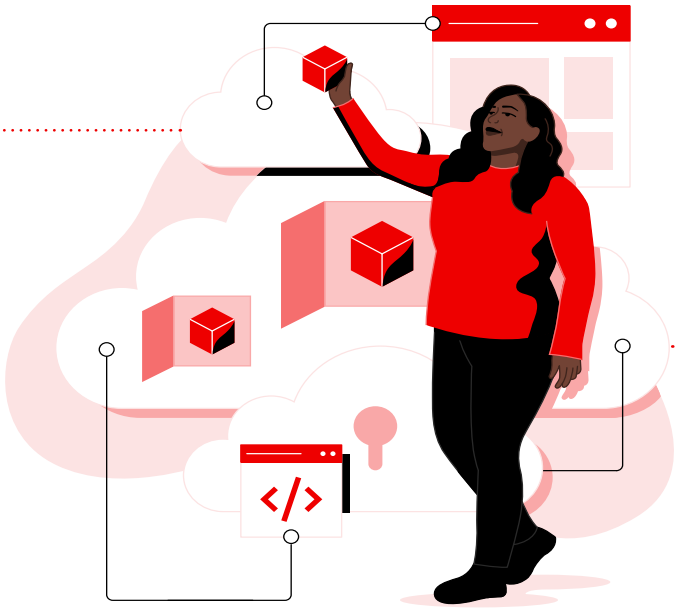
Grâce à des partenariats avec NVIDIA, AMD, Intel et d'autres partenaires technologiques pour l'IA, la solution Red Hat AI Enterprise fournit une plateforme d'IA d'entreprise de bout en bout qui peut évoluer dans le cloud hybride, avec des déploiements plus rapides et une meilleure efficacité. Les programmes de validation et de certification de Red Hat garantissent la pleine exploitation du matériel et une gestion optimisée des charges de travail pour une utilisation efficace des GPU ainsi que des performances maximales pour les clients.

La présence de Red Hat dans l'écosystème Hugging Face et dans le catalogue de serveurs MCP (Model Context Protocol) permet aux clients d'accéder à une bibliothèque sans cesse enrichie de modèles validés et d'outils préintégrés qui s'exécutent de manière cohérente avec la solution Red Hat AI Enterprise. En parallèle, les partenariats avec plusieurs fournisseurs d'accélérateurs aident les entreprises à tirer parti des GPU et du matériel spécialisé dans les environnements hybrides. Vous pouvez choisir en toute confiance les solutions de partenaires, technologies, modèles et outils les plus adaptés à vos besoins, avec la certitude qu'ils fonctionneront ensemble de manière fiable et que vous pourrez bénéficier de services spécialisés, d'une assistance et de formations pour vous aider à créer et mettre à l'échelle efficacement vos workflows d'IA.

Principaux partenaires

NVIDIA est un leader mondial des technologies d'IA, qui propose une architecture et une plateforme GPU de pointe pour l'IA et le calcul accéléré. Il offre d'excellentes performances et un haut niveau d'efficacité énergétique pour l'entraînement et le déploiement de grands modèles d'IA. NVIDIA et Red Hat ont noué un partenariat stratégique pour proposer une infrastructure et des logiciels cloud-native optimisés pour l'IA. En associant Red Hat AI Enterprise aux solutions NVIDIA, les clients bénéficient d'une suite complète, cloud-native et optimisée de logiciels d'IA et d'analyses de données, ainsi que du matériel nécessaire pour les exécuter. Les systèmes Red Hat AI Enterprise et NVIDIA DGX facilitent la gestion informatique de l'infrastructure d'IA. La solution NVIDIA GPU Operator automatise la gestion de tous les composants logiciels NVIDIA qui interviennent dans le provisionnement des GPU.

AMD et Red Hat collaborent pour offrir plus de choix en matière d'IA et de virtualisation dans l'ensemble des environnements de cloud hybride. Cette collaboration permet l'utilisation des processeurs EPYC AMD et des accélérateurs Instinct AMD avec la plateforme Red Hat AI Enterprise. Les entreprises peuvent ainsi exécuter des charges de travail d'entraînement, de réglage et d'inférence d'IA sur du matériel performant et rentable. Les efforts conjoints en matière d'ingénierie et de certification assurent la cohérence entre les systèmes AMD et la solution Red Hat AI Enterprise, avec notamment la prise en charge des images de conteneurs optimisées et des opérateurs Kubernetes, ainsi que la validation des performances des pipelines d'IA accélérés par GPU.



Intel et Red Hat collaborent pour offrir une infrastructure logicielle et des plateformes standards qui favorisent l'agilité et la flexibilité des datacenters. La distribution Intel du kit d'outils OpenVINO optimise et convertit les modèles d'IA en moteurs d'inférence hautes performances qui peuvent s'adapter automatiquement à des milliers de nœuds sur Red Hat AI Enterprise.

Témoignages de réussite

Turkish Airlines

La compagnie aérienne Turkish Airlines utilise Red Hat AI pour moderniser ses processus d'exploitation et ouvrir la voie à l'innovation basée sur l'IA dans l'aviation. Grâce à la standardisation de son environnement sur une plateforme d'IA ouverte et évolutive, la compagnie aérienne peut accélérer le développement des modèles, améliorer les services aux passagers et rationaliser la prise de décisions opérationnelles, démontrant ainsi que l'IA hybride peut transformer l'un des plus grands réseaux aériens au monde.

En savoir plus

Denizbank

Denizbank utilise Red Hat AI pour accélérer l'innovation en matière d'IA au sein de son écosystème bancaire numérique. Grâce à la modernisation de son infrastructure d'IA basée sur une plateforme ouverte et évolutive, la banque peut accélérer les expérimentations, améliorer la fiabilité des modèles et proposer des expériences client plus intelligentes, apportant ainsi la preuve que l'IA hybride aide les institutions financières à évoluer plus vite tout en préservant la rigueur de la posture de sécurité et de la gouvernance.

En savoir plus

AGESIC

L'AGESIC, l'agence gouvernementale numérique de l'Uruguay, utilise Red Hat AI pour standardiser et mettre à l'échelle l'IA dans plus de 180 entités publiques. La plateforme d'IA hybride prend en charge les pratiques MLOps, renforce la sécurité et aide les équipes à créer, déployer et contrôler des applications d'IA qui améliorent les services proposés aux citoyens.

En savoir plus

Envie d'exploiter tout le potentiel de vos données ?

L'IA transforme presque tous les aspects d'une entreprise. Chez Red Hat, nous pouvons vous aider à créer un environnement d'IA prêt pour la production qui accélère le développement et la distribution des applications intelligentes afin de soutenir vos objectifs métier.

Découvrez comment Red Hat AI Enterprise peut vous aider à créer une plateforme unifiée pour l'IA.