Red Hat OpenShift AI

Une plateforme ouverte pour les modèles d'IA dans le cloud hybride

Points clés

Exploitez et mettez à l'échelle l'inférence d'IA et l'IA agentique à l'aide d'une plateforme d'applications à l'efficacité prouvée.

Améliorez l'efficacité opérationnelle de l'IA/AA grâce à une expérience utilisateur cohérente qui donne de l'autonomie aux équipes de science des données, d'ingénierie des données, de développement d'applications et DevOps.

Gagnez en flexibilité dans le cloud hybride en procédant à la création, à l'entraînement, au déploiement et à la surveillance des charges de travail d'IA/AA sur site, dans le cloud ou à la périphérie du réseau.

Passez aux applications intelligentes et à l'IA générative

Les technologies d'intelligence artificielle (IA), d'apprentissage automatique (AA) et d'apprentissage profond influent de façon considérable sur la stratégie de modernisation des applications dans une grande variété de secteurs d'activité et d'entreprises. La nécessité d'innover et d'exploiter les données pour obtenir une valeur stratégique et de nouvelles informations contribue à la hausse de l'utilisation des applications cloud-native basées sur l'IA ainsi que des méthodes MLOps et GenAIOps. Dans cette situation inédite, les défis peuvent s'avérer complexes : augmentation rapide des coûts pendant la mise en production, personnalisation difficile, contraintes de déploiement rigides ou réalisation des tâches nécessaires pour suivre le rythme de l'innovation. Les entreprises ont besoin de solutions qui réduisent les coûts d'inférence, simplifient la mise à l'échelle et la surveillance, et s'adaptent au changement constant.

La gamme Red Hat® AI accélère le développement et le déploiement de solutions d'IA pour les entreprises dans des environnements de cloud hybride. Avec ses fonctionnalités MLOps et GenAIOps, elle fait office de plateforme complète de gestion pour l'intégralité du cycle de vie de l'IA/AA. Cette offre s'articule autour de quatre axes stratégiques :

- Augmentation de l'efficacité grâce à un processus d'inférence rapide, flexible et efficace
- Simplification de l'expérience de connexion des modèles aux données
- Accélération de l'innovation pour l'IA agentique
- Mise à l'échelle de l'IA dans le cloud hybride de manière flexible et cohérente

Basée sur la plateforme d'applications de cloud hybride leader sur le marché Red Hat OpenShift, la solution Red Hat OpenShift® AI est le produit phare de la gamme Red Hat AI. Cette plateforme d'IA offre aux équipes d'ingénierie, de science des données et de développement une base d'IA/AA puissante pour créer et déployer des modèles génératifs et prédictifs ainsi que des applications basées sur l'IA à grande échelle. Elle permet aux entreprises de tester des modèles avec un vaste choix d'outils, de collaborer plus facilement et d'écourter le délai de mise sur le marché en utilisant une plateforme unique. Red Hat OpenShift AI associe l'environnement en libre-service dont ont besoin les équipes de science des données et de développement à la fiabilité indispensable aux équipes informatiques.

Accélérez le développement, l'entraînement, les tests et le déploiement

Conçue autour de technologies Open Source, Red Hat OpenShift AI est une plateforme MLOps flexible et évolutive qui offre aux équipes des fonctionnalités fiables et cohérentes pour faire des expériences, déployer des modèles et distribuer des applications innovantes. Elle accélère la distribution d'applications basées sur l'IA pour permettre aux entreprises de passer plus rapidement d'un projet pilote à des déploiements robustes, avec une plus grande maîtrise.

La plateforme propose une interface utilisateur intégrée ainsi que des outils de création, d'entraînement, de réglage, de déploiement et de surveillance pour les modèles d'IA prédictive et générative. Elle permet de déployer des modèles dans des environnements de cloud hybride et assure le niveau de contrôle et de protection nécessaire pour l'IA souveraine et privée. Cette approche garantit que les données sensibles et les modèles d'IA restent dans les limites géographiques ou organisationnelles définies, dans le respect des exigences strictes en matière de réglementation et de conformité.

Red Hat
OpenShift AI

Prévision d'un analyste concernant l'IA générative

« On estime que l'IA va fortement influer sur les budgets alloués aux infrastructures numériques en 2026, car les entreprises doivent choisir leur infrastructure en fonction de leurs exigences en matière de charges de travail et de données. 90 % des décisionnaires pensent que l'IA sera un élément central de leur budget consacré à l'infrastructure numérique et de leurs choix technologiques[1]. »

Simplifiez l'adoption de l'IA

Solution complémentaire de Red Hat OpenShift, OpenShift AI fournit une plateforme conçue pour favoriser l'adoption de l'IA et renforcer la confiance dans ce type d'initiative en associant les communautés Open Source et un écosystème d'IA robuste. Les entreprises sont ainsi plus libres de choisir la technologie d'IA/AA la mieux adaptée à leurs besoins. Les utilisateurs ont la possibilité de créer leurs propres modèles prédictifs ou de commencer avec un modèle d'IA générative externe, puis de l'améliorer à l'aide de la génération augmentée de récupération (RAG) sur l'un des nombreux serveurs de modèles disponibles sur la plateforme. Cette dernière offre un accès rapide à des modèles tiers optimisés et validés, tels que Llama, Mistral, DeepSeek et Granite, qui s'exécutent efficacement sur vLLM, disponible dans le référentiel Red Hat AI sur Hugging Face. Le catalogue permet aux utilisateurs de découvrir ces modèles et d'ajouter les leurs.

Améliorez la cohérence opérationnelle entre les équipes

Red Hat OpenShift AI offre une expérience utilisateur cohérente qui favorise la collaboration entre les équipes de science des données, d'ingénierie de l'IA, de développement et DevOps et donc le respect des calendriers de production des solutions d'IA. Cette plateforme permet d'accéder en libre-service aux workflows collaboratifs, de bénéficier de l'accélération par processeur graphique (GPU) et de rationaliser l'exploitation, pour une distribution cohérente des solutions d'IA à grande échelle dans les environnements de cloud hybride et à la périphérie du réseau.

L'équipe d'exploitation informatique bénéficie de configurations simplifiées et de workflows davantage automatisés sur une plateforme éprouvée qui peut évoluer en toute simplicité, tout en améliorant la gouvernance et la sécurité.

Renforcez la flexibilité dans le cloud hybride

Red Hat OpenShift AI permet l'entraînement, le déploiement et la surveillance des charges de travail d'IA/AA dans divers environnements (datacenters sur site, clouds ou systèmes air gap) afin de répondre aux exigences en matière de réglementation, de sécurité et de données. La plateforme est compatible avec de nombreux accélérateurs d'IA de fournisseurs tels que NVIDIA, AMD et Intel. Cette capacité peut être étendue pour créer un environnement GPU-as-a-Service qui permet aux entreprises de centraliser la gestion, le partitionnement et la planification des ressources de GPU, tout en offrant une observabilité détaillée de leur utilisation.

Exploitez l'IA générative et l'IA agentique

Il existe des interfaces spécialisées pour les projets d'IA générative, par exemple AI Hub (version préliminaire pour les développeurs), un tableau de bord destiné aux équipes d'ingénierie de plateforme qui rassemble un catalogue, un registre et des déploiements de modèles pour configurer et déployer des modèles et des serveurs MCP. Gen AI Studio (version préliminaire pour les développeurs) fournit des points de terminaison de ressources d'IA ainsi qu'un environnement d'expérimentation où les équipes d'ingénierie de l'IA et de développement d'applications peuvent accéder à des modèles déployés et à des serveurs MCP, puis les tester, les comparer et les évaluer.

OpenShift AI accélère l'IA agentique en fournissant une couche d'API unifiée ainsi qu'une base flexible et évolutive. La prise en charge de l'API Llama Stack et du MCP (version préliminaire) comprend une mise en œuvre en entreprise de l'API Llama Stack, offrant un point d'entrée unique et standardisé pour diverses fonctionnalités d'IA.

Parmi d'autres outils figurent l'évaluation des grands modèles de langage (LM Eval) et le test de leurs performances pour faciliter les déploiements de processus d'inférence en conditions réelles. LLM compressor fournit des algorithmes qui permettent de réduire la taille des modèles personnalisés d'une entreprise à l'aide de méthodes comparables à celles utilisées par Red Hat pour créer des modèles validés et optimisés dans le référentiel Red Hat AI sur Hugging Face.

1 *IDC Tech Supplier, « AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025 », document n° US53418426, octobre 2025 (Nécessite une connexion client)*

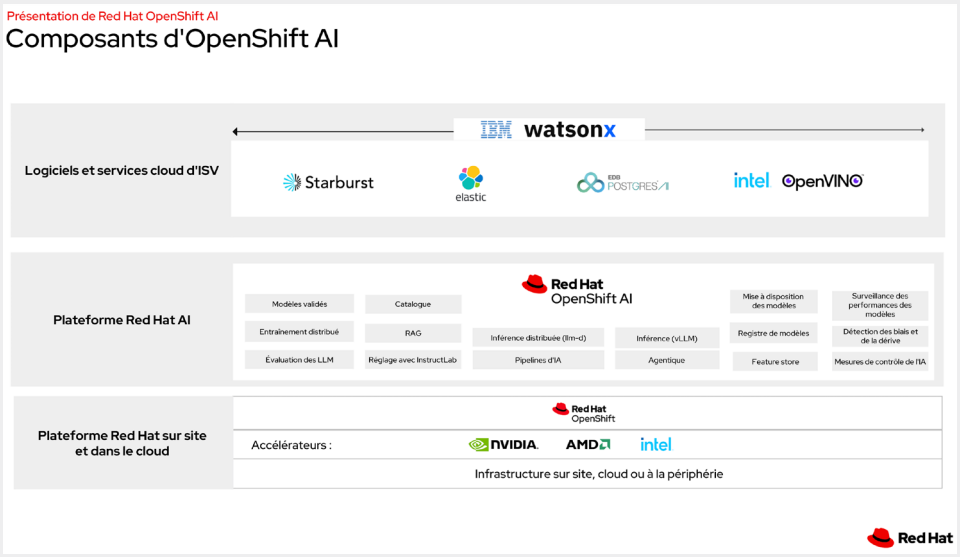

Figure 1 : Red Hat OpenShift AI et composants facultatifs de nos partenaires

Plusieurs autres fonctionnalités et outils essentiels compris dans Red Hat OpenShift AI forment une base solide :

- **Création et personnalisation de modèles.** Les data scientists peuvent réaliser des tests dans une interface utilisateur JupyterLab qui propose des images de notebooks sécurisées et prêtes à l'emploi avec des bibliothèques Pythons communes. Pour les projets d'IA générative, OpenShift AI permet de réaliser la RAG ainsi que l'entraînement distribué d'InstructLab, en fournissant des outils d'alignement de modèle pour contribuer plus efficacement aux modèles d'IA générative.
- **Mise à disposition des modèles.** Red Hat OpenShift AI propose différents frameworks à l'aide de KServe comme moteur principal pour gérer la mise à disposition des modèles afin de simplifier le déploiement de l'apprentissage automatique prédictif ou des modèles de fondation dans les environnements de production. Pour les LLM qui nécessitent une évolutivité maximale, OpenShift AI permet la mise à disposition des modèles en parallèle avec des environnements d'exécution vLLM. Le framework llm-d permet d'optimiser l'inférence des LLM en divisant le pipeline en services modulaires, ce qui rend possibles la mise à l'échelle automatique intelligente et le routage efficace des requêtes.
- **Pipelines d'IA.** Red Hat OpenShift AI inclut un composant qui permet d'organiser les tâches d'IA sous forme de pipelines et d'en construire à l'aide d'une interface graphique. Les entreprises peuvent combiner des processus tels que la préparation des données, la création et la mise à disposition des modèles.
- **Surveillance des modèles.** Red Hat OpenShift AI aide les spécialistes de l'exploitation à surveiller les indicateurs d'exploitation et de performances des serveurs de modèles et des modèles déployés. Les utilisateurs peuvent visualiser ces indicateurs dans des formats prédéfinis ou intégrer les données à d'autres services d'observabilité.
- **Distribution des charges de travail.** Grâce à la distribution des charges de travail, les équipes peuvent accélérer le traitement des données ainsi que l'entraînement, le réglage et le déploiement des modèles. Cette fonctionnalité aide à définir les priorités et la répartition des tâches ainsi qu'à utiliser les nœuds de manière optimale. Le haut niveau de compatibilité des GPU permet de gérer les exigences des charges de travail liées aux modèles de fondation.

- **Garde-fous de l'IA, détection des biais et de la dérive.** Red Hat OpenShift AI fournit des outils qui aident les équipes de science des données et d'ingénierie de l'IA à évaluer la justesse et l'objectivité des modèles en fonction des données d'entraînement, mais aussi après leur déploiement, dans des cas d'utilisation concrets. Les mécanismes de contrôle de l'IA offrent un cadre personnalisable qui met en œuvre des mesures de sécurité essentielles et garantit que les modèles sont fiables, transparents et équitables pour une utilisation en production. Les outils de détection de la dérive incluent l'analyse de la distribution des données d'entrée des modèles d'AA déployés, ce qui permet de repérer tout écart significatif entre les données d'entraînement et celles utilisées lors des opérations d'inférence.
- **Catalogue et registre.** Red Hat OpenShift AI fournit un catalogue de modèles interne ainsi qu'un catalogue personnalisé dans lesquels les équipes d'ingénierie de plateforme peuvent découvrir, comparer et évaluer des modèles d'IA générative optimisés. Cette plateforme propose également un registre centralisé permettant aux équipes de science des données et d'ingénierie de l'IA de partager, modifier, déployer et suivre les modèles d'IA prédictive et générative, les métadonnées et les artéfacts des modèles.
- **Feature store.** Ce référentiel permet de gérer des variables propres et bien définies pour les modèles d'AA, afin d'améliorer les performances et d'accélérer les workflows.

Outils pour le cycle de vie complet de l'IA

Red Hat OpenShift fournit aux entreprises les services et logiciels nécessaires pour l'entraînement, le déploiement et la mise en production de leurs modèles (voir la Figure 2).

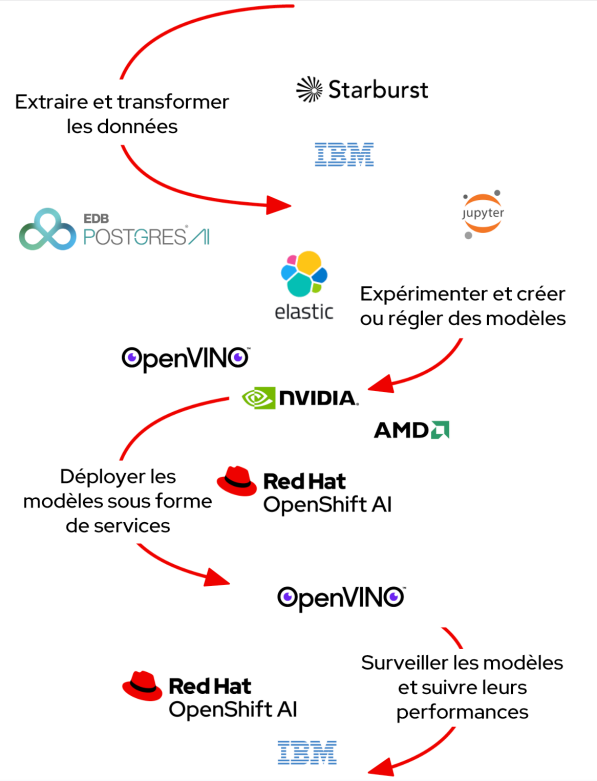

Figure 2 : cycle de vie standard de mise en œuvre d'un modèle

Le tableau de bord de Red Hat OpenShift AI centralise l'accès à l'ensemble des applications et supports de documentation pour faciliter la prise en main. Des guides de démarrage, directement disponibles via le tableau de bord, délivrent les meilleures pratiques concernant les composants et les logiciels partenaires intégrés les plus utilisés. Les data scientists peuvent ainsi se former et commencer à travailler plus vite. Les sections suivantes décrivent les outils des partenaires technologiques intégrés à Red Hat OpenShift AI. Certains de ces outils nécessitent une licence supplémentaire à obtenir auprès du partenaire concerné.

Starburst

Starburst accélère les analyses grâce à une solution simple et rapide qui permet de tirer parti des données pour améliorer le fonctionnement de l'entreprise. Disponible en tant que logiciel autogéré ou service entièrement géré, Starburst démocratise l'accès aux données et fournit des informations complètes aux utilisateurs. Cette plateforme est basée sur l'outil Open Source Trino (anciennement PrestoSQL), une référence dans le domaine des moteurs Structured Query Language (SQL) à traitement massivement parallèle (MPP). Conçue et gérée par les spécialistes de Trino, elle permet d'interroger des ensembles de données très variés, dans tous types d'environnements, sans les déplacer.

Compatible avec les services évolutifs de calcul et de stockage dans le cloud de Red Hat OpenShift, Starburst offre de bonnes performances en matière de stabilité, de sécurité, d'efficacité et de rentabilité. Les avantages offerts sont nombreux :

- **Automatisation.** Avec Starburst et les opérateurs Red Hat OpenShift, les clusters sont capables de s'autoconfigurer, de s'autorégler et de s'autogérer.
- **Haute disponibilité et arrêt graduel.** L'équilibreur de charge de Red Hat OpenShift peut maintenir actifs des services comme le coordinateur Trino.
- **Évolutivité.** Red Hat OpenShift met automatiquement à l'échelle le cluster de calcul Trino en fonction de la charge d'interrogation.

HPE Machine Learning Data Management Software

Les entreprises ont besoin de solutions de gestion des données qui facilitent l'ensemble du processus, des expérimentations menées individuellement aux déploiements majeurs dans l'entreprise. Grâce au logiciel HPE Machine Learning Data Management Software (anciennement Pachyderm), les data scientists peuvent créer et mettre à l'échelle des pipelines d'AA conteneurisés et basés sur des données. De plus, un système de gestion automatique des versions garantit la traçabilité des données. Spécialement conçu pour résoudre les problèmes concrets liés à la science des données, ce logiciel offre les bases nécessaires pour automatiser et mettre à l'échelle le cycle de vie de l'AA de manière reproductible. Il convient pour de nombreux cas d'utilisation, tels que les données non structurées, les entrepôts de données, le traitement du langage naturel, les processus ETL (extraction, transformation et chargement) pour les vidéos et les images, les services financiers et les sciences de la vie. Il offre les capacités suivantes :

- Gestion automatique des versions pour un suivi hautes performances des modifications apportées aux données
- Pipelines conteneurisés basés sur des données qui accélèrent le traitement des données et réduisent les coûts de calcul
- Processus de traçabilité des données immuable permettant l'enregistrement fixe des activités et ressources du cycle d'AA
- Console avec une interface intuitive de visualisation des graphes orientés acycliques qui simplifie le débogage et assure la reproductibilité
- Prise en charge des notebooks Jupyter via le module JupyterLab Mount Extension qui permet d'accéder en quelques clics aux versions de données.

- Outils robustes pour le déploiement et l'administration du logiciel HPE Machine Learning Data Management Software au sein de différentes équipes dans l'entreprise

NVIDIA accélère le déploiement de solutions d'IA

Parce que les applications d'IA/AA jouent un rôle de plus en plus important dans leur réussite, les entreprises ont besoin de plateformes qui peuvent gérer des charges de travail complexes, optimiser l'utilisation du matériel et garantir une bonne évolutivité. Le traitement évolutif des données, l'analyse des données, l'entraînement des modèles d'AA et l'inférence sont des tâches de calcul qui consomment énormément de ressources. NVIDIA propose des logiciels qui permettent d'accélérer tous les aspects de la science des données grâce aux capacités de traitement parallèle des GPU.

NVIDIA NIM améliore la gestion et les performances des GPU NVIDIA dans l'environnement Red Hat OpenShift, en permettant aux applications d'IA d'exploiter tout le potentiel du matériel et des logiciels d'IA de NVIDIA. L'intégration de NVIDIA NIM et de Red Hat OpenShift AI permet d'améliorer l'allocation des ressources, l'efficacité et la vitesse d'exécution des charges de travail d'IA.

Kit d'outils OpenVINO d'Intel

Le kit d'outils OpenVINO d'Intel accélère le développement et le déploiement des applications d'inférence d'apprentissage profond hautes performances sur les plateformes Intel. Il permet d'adopter, d'optimiser et de régler presque tous les modèles de réseaux de neurones ainsi que d'exécuter des opérations d'inférence d'IA complètes grâce à l'écosystème OpenVINO d'outils de développement.

- **Modélisation :** les équipes de développement peuvent utiliser leurs propres modèles d'apprentissage profond. Pour accélérer la mise sur le marché, elles peuvent également utiliser des modèles préentraînés et préoptimisés, disponibles dans le kit d'outils OpenVINO de Hugging Face et d'Intel. OpenVINO prend en charge Pytorch, ONNX, TensorFlow et d'autres formats de modèle courants.
- **Optimisation :** le kit d'outils OpenVINO propose plusieurs manières plus pratiques et performantes de convertir des modèles, ce qui permet aux équipes de développement d'exécuter des modèles d'IA plus rapidement et efficacement. Les modèles aux formats PyTorch, ONNX, TensorFlow, TensorFlow Lite, JAX ou PaddlePaddle n'ont pas besoin d'être convertis et peuvent être utilisés tels quels pour l'inférence. La conversion au format OpenVINO IR permet d'obtenir des performances optimales, qui peuvent être encore améliorées à l'aide des fonctions de compression et de quantification des pondérations disponibles dans le framework NNCF (Neural Network Compression Framework) d'OpenVINO. Ces mêmes fonctions réduisent également l'empreinte de stockage et d'exécution.
- **Déploiement :** l'API OpenVINO Runtime Inference Engine est conçue pour être intégrée aux applications afin d'accélérer le processus d'inférence. Elle suit une approche de type « coder une fois, déployer partout » qui permet d'exécuter efficacement les tâches d'inférence sur du matériel Intel varié, notamment des processeurs, des GPU, des NPU et des FGPA. La bibliothèque d'extensions OpenVINO GenAI simplifie le déploiement des charges de travail d'IA générative et permet, dans de nombreux cas, de limiter le code nécessaire à trois ou cinq lignes. OpenVINO Model Server offre plusieurs fonctions pour les scénarios agentiques et de mise à disposition des modèles, ce qui réduit encore davantage les efforts de développement.

EDB

Puissante et intelligente, la plateforme EDB Postgres AI gère les charges de travail transactionnelles, analytiques et d'IA. Elle offre un niveau de flexibilité inégalé, que les données soient stockées sur site ou dans le cloud. En tant que leader mondial des solutions de bases de données Postgres pour les entreprises, EDB fournit une plateforme d'IA et de données souveraine, ouverte et adaptée aux entreprises qui permet de passer les projets d'IA en production jusqu'à trois fois plus rapidement. Intégrée à Red Hat OpenShift AI, la solution EDB Postgres AI permet de créer des bases de connaissances d'IA robustes pour la RAG, unifiant ainsi les données, les modèles et les applications d'IA au sein d'une plateforme d'IA souveraine, complète et déployable dans tous les environnements. Cette transformation des données d'exploitation essentielles en ressources optimisées pour l'IA

permet d'obtenir jusqu'à 30 % d'efficacité supplémentaire et de simplifier l'utilisation des données privées, y compris les données non structurées, pour alimenter les résultats des modèles dans la base de connaissances de l'entreprise.

Elastic

La solution Elastic Search AI Platform (basée sur ELK Stack[2]) associe la précision de la recherche et l'intelligence de l'IA pour permettre aux utilisateurs de créer des prototypes et d'intégrer plus rapidement des LLM, ainsi que d'utiliser l'IA générative pour créer des applications rentables et évolutives. Grâce à cette solution, les utilisateurs peuvent créer des applications de RAG innovantes, résoudre proactivement des problèmes d'observabilité et éliminer les menaces complexes. Le moteur Elasticsearch peut être déployé dans les environnements où se trouvent les applications : sur site, sur une plateforme de cloud public ou même dans un environnement air-gap.

Elastic s'intègre aux modèles de l'écosystème, tels que Red Hat OpenShift AI, Hugging Face, Cohere et OpenAI, via un simple appel d'API. Cette approche garantit un code propre pour gérer l'inférence hybride des charges de travail de RAG, avec les fonctions suivantes :

- Découpage, connecteurs et indexeurs pour ingérer différents ensembles de données dans la couche de recherche
- Recherche sémantique avec Elastic Learned Sparse EncodeR (ELSER), le modèle d'AA intégré, et le modèle d'intégration E5, pour une recherche vectorielle multilingue
- Documentation et sécurité sur le terrain, avec mise en œuvre des autorisations et des droits qui correspondent au contrôle d'accès basé sur les rôles de l'entreprise

Avec la solution Elastic Search AI Platform, vous entrez dans une communauté internationale de développeurs qui aiment chercher des solutions et apporter leur aide. Retrouvez la communauté Elastic sur Slack, sur nos réseaux sociaux ou dans nos forums de discussion.

Conclusion

Grâce à la plateforme Red Hat OpenShift AI, les entreprises peuvent expérimenter des modèles plus facilement, collaborer davantage et, en fin de compte, accélérer le déploiement de leurs applications basées sur l'IA. Elle permet aux équipes de science des données et d'ingénierie de l'IA de créer et déployer des modèles dans le cloud hybride. Les équipes d'exploitation informatique et d'ingénierie de plateforme profitent de ses capacités MLOps et GenAIOps pour déployer les modèles plus rapidement en production. Grâce à l'accès en libre-service, notamment aux GPU, les équipes de développement, d'ingénierie de l'IA et de science des données peuvent innover sur une plateforme d'applications que le service informatique utilise déjà en toute confiance. Red Hat OpenShift AI fournit une plateforme fiable, complète et cohérente qui propose des facteurs de différenciation uniques en matière de performance de l'inférence, d'IA agentique et d'exploitation évolutive du cloud hybride, le tout avec un écosystème robuste de partenaires.

En savoir plus

Lancez-vous dès aujourd'hui en consultant la page consacrée à Red Hat OpenShift AI.

À propos de Red Hat

Red Hat aide ses clients à standardiser leurs environnements, à développer des applications cloud-native et à intégrer, automatiser, sécuriser et gérer des environnements complexes en offrant des services d'assistance, de formation et de consulting primés.

2 *La pile ELK comprend Elasticsearch, Kibana, Beats et Logstash.*

facebook.com/redhatinc
@RedHatFrance
linkedin.com/company/red-hat

EUROPE, MOYEN-ORIENT ET AFRIQUE (EMEA)
00800 7334 2835
europe@redhat.com

FRANCE
00 33 1 41 91 23 23
fr.redhat.com

fr.redhat.com
#2876428_1125