Accélérer l'adoption de l'IA pour les services financiers avec Red Hat

Réduction du délai de mise sur le marché pour les solutions d'IA/AA avec une plateforme de bout en bout

Des modèles d'IA plus complexes, une adoption plus difficile

Les institutions financières cherchent à saisir les opportunités qu'offre l'adoption de l'intelligence artificielle (IA). Les technologies d'apprentissage profond, d'IA conversationnelle et générative évoluent à une vitesse telle que les solutions d'IA peuvent aujourd'hui être utilisées dans de nombreux domaines. En parallèle, l'accroissement de la complexité des modèles exacerbe les défis liés à l'exécution et en crée d'autres. Voici quelques exemples de ces défis :

- **Processus de développement autonome :** le développement et l'entraînement des modèles d'IA et d'apprentissage automatique (AA) sont actuellement en grande partie réalisés dans des environnements dédiés et nécessitent des ressources spéciales, comme des processeurs graphiques (GPU). Le provisionnement d'environnements d'IA/AA prend beaucoup de temps et peut bloquer le déploiement de nouveaux services basés sur l'IA.
- **Mise à l'échelle, flexibilité et optimisation des ressources :** les solutions d'IA/AA nécessitent des composants aux besoins en ressources variés, comme des unités centrales de traitement (CPU), de la mémoire, de l'espace disque et du matériel spécialisé (GPU), des unités de traitement de tenseur (TPU) et des circuits FPGA (Field-Programmable Gate Array). La mise à l'échelle de ces solutions doit souvent être effectuée dans le cadre d'une approche de cloud hybride.
- **Surveillance et écarts :** les modèles d'IA/AA requièrent une surveillance continue et des mises à jour régulières pour détecter et corriger les écarts. La solution Red Hat® OpenShift® facilite l'intégration continue des mises à jour des modèles en fournissant une infrastructure de surveillance reposant sur des normes qui peut connecter une application de surveillance des écarts au pipeline de développement de l'IA/AA.
- **Sécurité de la chaîne d'approvisionnement des modèles :** l'écosystème d'outils de développement de l'IA/AA repose essentiellement sur des frameworks Open Source et communautaires. Dans cet environnement, il devient de plus en plus difficile de garantir l'hygiène de la chaîne d'approvisionnement des logiciels. Les équipes de développement veulent pouvoir utiliser les outils les plus récents, mais les entreprises doivent vérifier la sécurité de ces outils en s'assurant qu'ils ne contiennent aucun artéfact vulnérable ou malveillant.

Des avantages qui réduisent considérablement les complexités

Voici quelques avantages de la solution d'IA/AA proposée pour les institutions financières :

- Une plateforme de bout en bout pour le développement, l'entraînement et l'inférence des modèles, qui permet d'assurer la cohérence de l'exploitation dans les clouds publics et privés, ainsi que de limiter les points de friction entre les différentes phases du processus
- Des fonctionnalités en libre-service qui réduisent le délai de rentabilisation des environnements d'AA
- Un ensemble cohérent d'outils et de bibliothèques d'AA Open Source, associé à un vaste écosystème de technologies Open Source et prises en charge par nos partenaires
- Des processus de développement et de déploiement des modèles d'AA plus rapides, avec des fonctionnalités de surveillance et d'itération qui garantissent que les modèles déployés restent à jour

Étude de cas : grands modèles de langage

Afin d'illustrer les défis et les avantages pour les services financiers, prenons l'exemple de la mise en œuvre d'une solution reposant sur un grand modèle de langage (LLM), comme GPT-4, BLOOM, BART, DOLLY, etc. Ces types de solutions sont utilisés pour numériser des documents dans le cadre des processus d'intégration ou de connaissance du client, pour analyser des rapports regroupant des données environnementales, sociales et de gouvernance d'entreprise (ESG) ou pour mettre en œuvre des solutions conversationnelles, comme des chatbots.

Ces solutions s'appuient souvent sur de grands modèles d'AA contenant des centaines de millions, voire des milliards de paramètres. La création de ces modèles est complexe et demande beaucoup de travail et de puissance de calcul. Il est donc courant d'utiliser des modèles de fondation ou préentraînés comme base. Ces modèles étant généralement entraînés avec des ensembles de données génériques, leur application au contexte spécifique d'un cas d'utilisation des services financiers exige un entraînement supplémentaire propre au domaine ou à l'entreprise avec un ensemble réduit de données locales, par le biais d'un réglage fin ou d'un apprentissage par transfert. La figure 1 présente un exemple d'architecture pour ce type de solution.

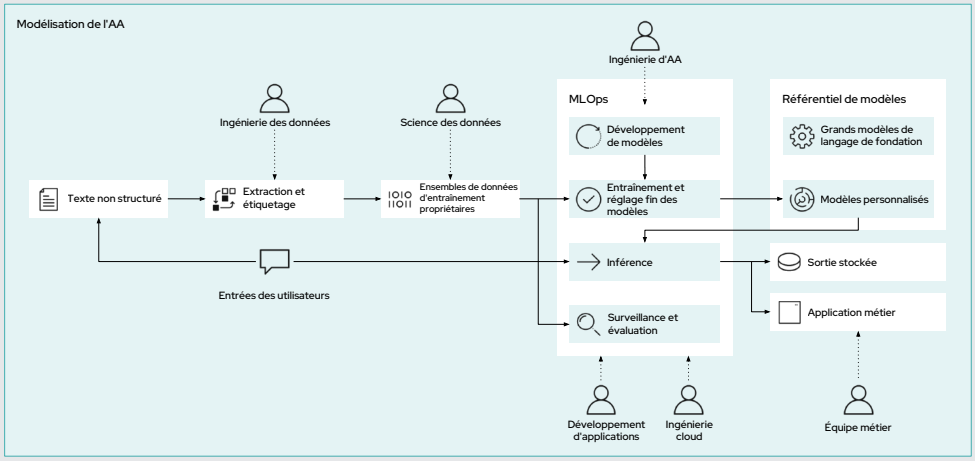

Figure 1 : aperçu de la modélisation de l'AA

Vue d'ensemble des capacités

Architecture de la solution

Nous proposons une plateforme qui héberge de manière efficace et productive le cycle de vie complet de l'IA/AA (développement, entraînement et inférence). Cette technologie s'exécute sur les principales formes d'infrastructure, de la virtualisation bare metal et sur site aux clouds publics les plus utilisés, ce qui permet d'exploiter la même plateforme avec les mêmes outils et les mêmes processus MLOps.

Nous comprenons l'intérêt de l'Open Source et l'importance de la protection de la chaîne d'approvisionnement des logiciels. C'est pourquoi nous nous engageons auprès des communautés en amont pour développer de nouveaux logiciels de qualité tout en nouant des relations de confiance. Dans le cadre de cette démarche, nous sélectionnons, prenons en charge et certifions un large éventail d'outils en amont nécessaires aux équipes de développement de l'IA/AA. Laissez-nous nous charger de comprendre la chaîne d'approvisionnement en amont et vous proposer un produit sur lequel vous pouvez compter, avec une assistance continue.

Composants de la plateforme

Système d'exploitation

Notre architecture d'IA/AA repose sur Red Hat Enterprise Linux®, un système d'exploitation qui peut s'exécuter sur une infrastructure de déploiement moderne sur site ou dans un environnement cloud ou bare metal, ou encore sur des machines virtuelles. Il est certifié pour fonctionner avec les solutions d'un très vaste écosystème de matériel et les clouds des principaux fournisseurs, notamment Amazon Web Services (AWS), Google Cloud, IBM Cloud for Financial Services, Oracle Cloud et Microsoft Azure. La plateforme Linux améliore la sécurité, les performances et la prise en charge, et permet d'automatiser les processus efficacement avec Red Hat Ansible® Automation Platform. Enfin, Red Hat Enterprise Linux prend en charge le matériel spécialisé destiné au développement de modèles d'IA/AA, notamment les GPU et les circuits FPGA.

Orchestration des conteneurs

Tout comme les applications conçues sur mesure et celles qui sont disponibles à l'achat, la grande majorité des bibliothèques et outils Open Source utilisés dans les processus d'IA/AA sont conteneurisés. Les modèles d'AA de production ou préentraînés sont mis en paquets dans des images de conteneurs. De plus, les processus d'IA/AA impliquent des interactions entre divers composants qui doivent pouvoir être mis à l'échelle de manière flexible. Une plateforme élastique est nécessaire pour l'ensemble de ces composants, que ce soit pour l'entraînement de nouveaux modèles nécessitant beaucoup de ressources de calcul, les moteurs d'inférence à haut débit et les environnements de développement de modèles utilisés par les équipes de science des données. La principale solution pour le déploiement et l'orchestration des charges de travail conteneurisées est Red Hat OpenShift, une distribution de Kubernetes. C'est la plateforme la plus fréquemment utilisée pour les outils de développement d'IA Open Source et tiers. Elle garantit à vos équipes de développement l'accès aux frameworks d'IA/AA dont elles ont besoin pour réduire le délai de rentabilisation. Red Hat OpenShift fournit également des opérateurs qui permettent d'automatiser le déploiement de composants pour le libre-service tout en réduisant les coûts d'exploitation.

Stockage évolutif à la sécurité renforcée

Les projets d'IA/AA nécessitent de grandes quantités de données d'entraînement pour permettre la création de modèles précis. Ces données peuvent être historiques ou actuelles, issues de sources telles que les flux de données de marché, les appareils de l'Internet des objets (IoT) ou les systèmes d'observabilité. Dans tous les cas, leur stockage doit être intuitif et les équipes de développement doivent pouvoir y accéder à plusieurs reprises. Avec Red Hat OpenShift Data Foundation, qui repose sur Red Hat Ceph® Storage, le stockage logiciel Open Source est pris en charge et intégré. La solution de stockage logiciel OpenShift Data Foundation s'intègre à Red Hat OpenShift et permet une mise à l'échelle à moindre coût pour atteindre plusieurs pétaoctets et au-delà. Les données de diffusion en continu peuvent être consommées avec AMQ Streams, qui repose sur Apache Kafka, pour fournir aux équipes de développement un accès répété à ces données. Les solutions OpenShift Data Foundation et AMQ Streams, mises en paquets dans des conteneurs, peuvent être gérées à l'aide de Red Hat OpenShift afin que plusieurs équipes de développement puissent travailler dans un contexte de libre-service.

Fonctionnalités de la plateforme

Libre-service

Avec Red Hat OpenShift, les projets et les équipes de développement peuvent être intégrés à la demande et les ressources peuvent être adaptées à la hausse ou à la baisse selon les besoins. En outre, le matériel spécialisé coûteux, comme les GPU, peut être regroupé dans des pools et partagé. La conformité et la sécurité de la chaîne d'approvisionnement des logiciels sont intégrées à tous les niveaux.

Surveillance et observabilité avancées

Red Hat OpenShift utilise l'outil Open Source et standardisé Prometheus pour la surveillance et est aussi compatible avec des outils de surveillance tiers, comme Splunk. Les pipelines MLOps sont intégrés à une infrastructure flexible et centralisée qui gère leur surveillance et les alertes. Avec le suivi des performances des modèles, la mise à l'échelle peut être automatisée et des alertes sont émises lorsque le niveau de précision est faible.

Agilité

La modélisation de l'IA/AA est un processus itératif. Les équipes d'ingénierie des données et de science des données explorent les chemins tracés par les données. Le parcours de développement d'un modèle est semé d'obstacles, de feux verts, de stops, de grands boulevards et d'impasses. Ces équipes se retrouvent généralement devant deux difficultés : l'accès à des données de qualité provenant de différentes sources (bases de données, systèmes de fichiers, flux, API) et la conformité avec les obligations réglementaires et les normes de sécurité. Au niveau des outils, ces équipes doivent gérer les versions dans un grand nombre de bibliothèques ainsi que mettre à jour les outils existants et en adopter d'autres. Avec nos solutions, nous simplifions le pipeline d'IA/AA en offrant une expérience cohérente aux spécialistes dans un environnement de cloud hybride afin d'accélérer les projets d'IA/AA.

Par rapport au développement d'applications traditionnelles, le développement d'applications d'IA/AA exige de mettre à jour les applications elles-mêmes ou les modèles d'IA sur lesquels elles reposent. En plus de l'entraînement initial d'un modèle dans le cadre de l'AA, les techniques d'IA/AA permettent de continuellement mettre à jour le modèle. Les modèles présentent ainsi des avantages que les applications traditionnelles n'offrent pas, mais il faut régulièrement corriger et mettre à jour les modèles afin d'améliorer les performances. Avec Red Hat OpenShift, les équipes qui développent des applications peuvent adapter les composants de la chaîne d'outils MLOps de manière transparente. Lorsque le modèle de l'application doit être mis à jour, les ressources d'entraînement (plus coûteuses) avec des GPU et d'autres composants spécialisés peuvent être attribuées et étendues manuellement. Dès que la mise à jour est terminée, Red Hat OpenShift peut réattribuer ces ressources selon les besoins.

Évolutivité et élasticité pour l'entraînement et l'inférence

La phase d'entraînement de la modélisation de l'IA/AA est l'une des opérations les plus gourmandes en ressources du pipeline MLOps. C'est au cours de celle-ci que les instances d'outils d'IA à très grande échelle sont utilisées et que la demande en matériel spécialisé (GPU, TPU et circuits FPGA notamment) d'entreprises comme Nvidia est la plus élevée. Pour chaque équipe ou projet, il est préférable d'effectuer l'entraînement dans un environnement qui lui est propre. Grâce à son infrastructure partagée, notre architecture d'IA/AA offre des avantages considérables en matière d'efficacité et d'économies. Avec Red Hat OpenShift, les équipes de développement disposent d'un accès virtuel et à la demande à l'intégralité du cluster et n'ont pas besoin d'accumuler des ressources dédiées à un prix élevé. Kubernetes orchestre cet accès et sert d'intermédiaire pour garantir la répartition de ces ressources selon les besoins métier.

Écosystème ouvert

Notre plateforme d'IA/AA est entièrement Open Source, comme tous les produits Red Hat. L'écosystème Open Source d'outils et de technologies pour les spécialistes de l'IA/AA englobe :

- des bibliothèques d'AA ;
- la gestion du cycle de vie de l'IA/AA ;
- la gestion de l'accès aux données, de la qualité des données et des métadonnées ;
- la détection des biais et l'explicabilité ;
- des modèles préentraînés.

Parce que cet écosystème est ouvert et que la plateforme est flexible, ces outils peuvent être utilisés dans diverses combinaisons selon les besoins des solutions. En outre, l'utilisation d'une plateforme ouverte favorise l'innovation continue, car les technologies, outils et modèles émergents peuvent toujours être intégrés à la solution.

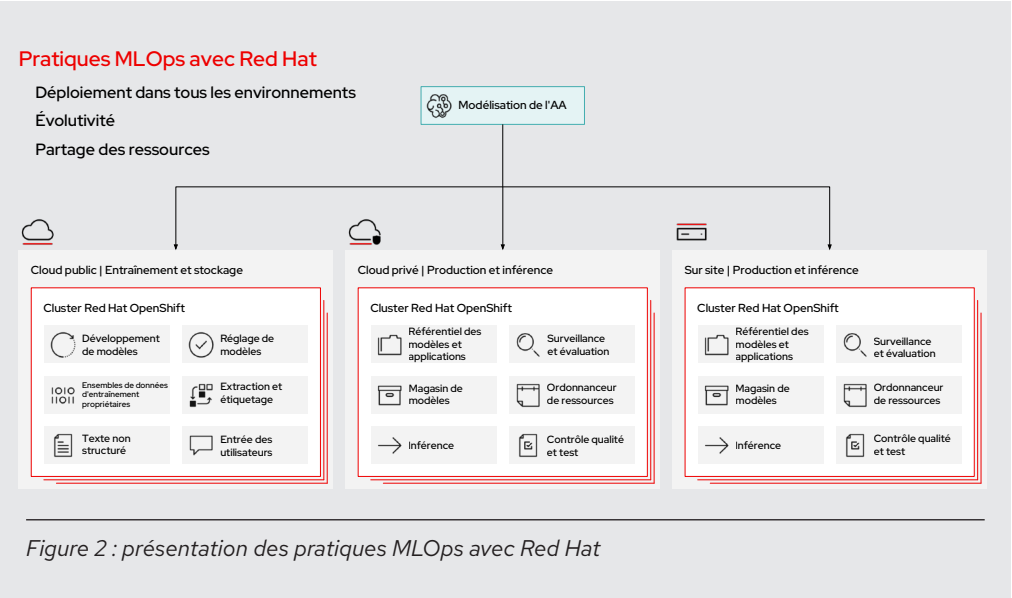

Figure 2 : présentation des pratiques MLOps avec Red Hat

À propos de Red Hat

Premier éditeur mondial de solutions Open Source, Red Hat s'appuie sur une approche communautaire pour fournir des technologies Linux, de cloud hybride, de conteneurs et Kubernetes fiables et performantes. Red Hat aide ses clients à développer des applications cloud-native, à intégrer des applications nouvelles et existantes ainsi qu'à gérer et automatiser des environnements complexes. Conseiller de confiance auprès des entreprises du Fortune 500, Red Hat propose des services d'assistance, de formation et de consulting reconnus qui apportent à tout secteur les avantages de l'innovation ouverte. Situé au cœur d'un réseau mondial d'entreprises, de partenaires et de communautés, Red Hat participe à la croissance et à la transformation des entreprises et les aide à se préparer à un avenir toujours plus numérique.

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
#420450_0723