Optimisez vos investissements en IA générative grâce à Dell AI Factory

Analyse coûts-avantages de la mise en œuvre d'une solution Dell AI Factory par rapport à AWS et Azure

L'IA générative (GenAI) peut aider les organisations de tous types et de toutes tailles à atteindre leurs objectifs métier, mais il peut s'avérer difficile de choisir une solution de taille adaptée qui réponde aux besoins en matière de performances, de budget et de sécurité. Même s'il peut sembler judicieux d'héberger de grands modèles de langage (LLM) dans le Cloud pour optimiser la flexibilité, s'engager à déployer l'IA en dehors votre datacenter peut entraîner des problèmes budgétaires au fil du temps et coûter plus cher aux entreprises à long terme. Dans une enquête réalisée en 2024, 46 % des dirigeants d'entreprise ont déclaré que le coût de la mise en œuvre de l'IA était une préoccupation, ce qui représente une hausse significative par rapport à seulement 3 % l'année précédente. Les entreprises commencent à interrompre ou à reporter les initiatives d'IA en raison des coûts, car elles encourent des « coûts liés aux tokens, des coûts supplémentaires inattendus et une prolifération de l'IA ».[1]

Pour aider les entreprises à comprendre le coût total du déploiement et de la gestion des charges applicatives d'IA générative, y compris le réglage précis et l'inférence des modèles, nous avons comparé les coûts approximatifs sur 4 ans d'une solution Dell™ AI Factory sur site exploitant le matériel PowerEdge™ R660 et PowerEdge XE9680 avec deux options de paiement (une solution CAPEX à modèle de paiement traditionnel et une solution Dell APEX Pay-Per-Use basée sur abonnement) avec les solutions Amazon Web Services (AWS) SageMaker et Microsoft Azure Machine Learning similaires. Selon nos calculs, les solutions Dell AI Factory étaient les plus rentables des solutions sur 4 ans, parmi celles que nous avons comparées. La solution d'abonnements Dell APEX a réduit les coûts de 71 % par rapport à AWS et de 60 % par rapport à Azure. La même solution sur site Dell AI Factory sans abonnement (modèle CAPEX) réduirait les coûts sur 4 ans de 71 % par rapport à la solution AWS et de 61 % par rapport à la solution Cloud Azure que nous avons facturée. Lisez ce qui suit pour découvrir comment choisir d'exécuter l'IA générative sur site avec une solution Dell AI Factory sur site peut aider votre entreprise à tirer le meilleur parti de votre investissement.

Tirez davantage de votre investissement avec une solution Dell AI Factory

Économisez jusqu'à 71 % par rapport à une solution AWS concurrente

Atteignez votre seuil de rentabilité en 1 an

Économisez jusqu'à 61 % par rapport à une solution Azure concurrente

Atteignez votre seuil de rentabilité en 18 mois

Coût total de possession sur 4 ans pour une solution sur site Dell AI Factory par rapport aux environnements AWS et Azure | *Une valeur plus faible est préférable*

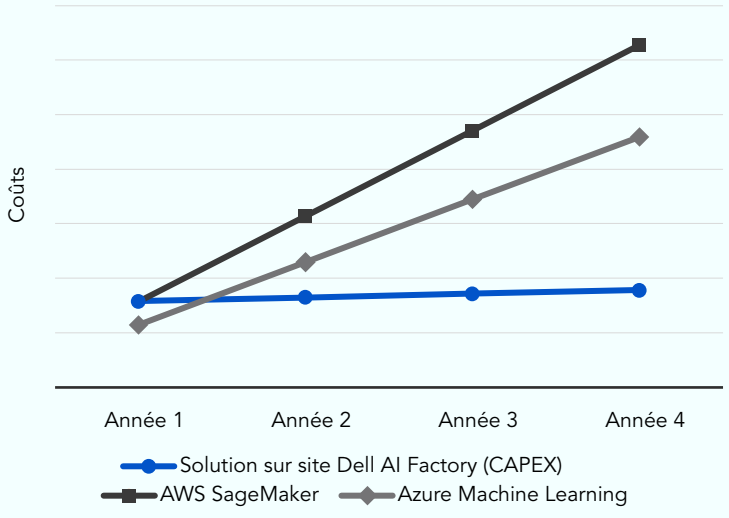


Figure 1 : Coûts relatifs d'une solution sur site Dell AI Factory (CAPEX) par rapport aux solutions AWS SageMaker et Azure Machine Learning sur 4 ans.

Concernant Dell AI Factory

Dell AI Factory est une approche complète conçue pour répondre à l'évolution des exigences des applications d'IA modernes. Optimisée par Dell, AI Factory combine une infrastructure, des logiciels et des services de pointe, offrant ainsi la flexibilité nécessaire pour accélérer les initiatives d'IA. Les organisations peuvent adopter l'intégralité de la solution, y compris tous les composants pour une approche fluide, ou personnaliser leur configuration en utilisant l'infrastructure et les services de Dell adaptés pour des résultats spécifiques d'IA générative. L'approche d'écosystème ouvert garantit que la compatibilité des solutions Dell AI Factory avec divers workflows et technologies.

Dell AI Factory offre une large gamme d'options d'infrastructure pour répondre aux besoins modernes des entreprises, y compris la possibilité de souscrire des abonnements Dell APEX, ce qui en fait un cadre précieux pour les entreprises cherchant à tirer parti de l'IA pour transformer leurs données en résultats commerciaux positifs. Pour en savoir plus sur les avantages de Dell AI Factory pour votre organisation, rendez-vous sur dell.com/ai.

Scénario de coût total de possession et présentation des solutions

De nouvelles charges applicatives impliquent souvent de nouveaux investissements. De nombreuses charges applicatives d'IA nécessitent des composants hautes performances en plus des grandes quantités de stockage contenant déjà vos données. Mettre en œuvre des charges applicatives d'IA implique d'équilibrer la sécurité, le temps, les performances et l'évolutivité, la facilité d'utilisation et les coûts. Pour vous donner une idée du coût des solutions à base d'IA, nous avons créé un scénario à l'aide du modèle Open source Llama 3 8B et comparé le coût d'exécution de la charge applicative dans quatre environnements différents. Notre scénario incluait quatre tâches spécifiques dans une charge applicative d'IA générative : codage et autre travail de développement de l'apprentissage automatique, tâches de traitement des données, tâches de réglage fin des modèles et tâches d'inférence. Ces tâches se combinent pour garder le modèle précis et à jour avec les dernières données générées par l'entreprise pour fournir des sorties de modèle optimales. Le Tableau 1 présente les spécifications générales des quatre environnements que nous avons étudiés. À noter : nous avons terminé toutes les recherches et les tarifications le 27 mars 2025. Les prix sont susceptibles d'être modifiés après cette date.

Tableau 1 : Détails de la solution pour la comparaison du coût total de possession.

Tâche	Serveur/Instance	Processeurs graphiques par serveur/instance	Autres achats
Solution sur site Dell AI Factory			
Gestion des clusters	3 serveurs PowerEdge R660	Sans objet	2x PowerSwitch S5232-ON Network Infrastructure et 1x PowerSwitch N3200-ON OOB Management
ordinateurs portables			
Traitement des données	2 serveurs PowerEdge XE9680 (avec 30 To de stockage chacun)	8x NVIDIA H100	
Réglage fin du modèle			
Inférence			
Solution d'abonnements sur site Dell APEX managée			
Gestion des clusters	3 serveurs PowerEdge R660	Sans objet	2x PowerSwitch S5232-ON Network Infrastructure et 1x PowerSwitch N3200-ON OOB Management
ordinateurs portables			
Traitement des données	2 serveurs PowerEdge XE9680 (avec 30 To de stockage chacun)	8x NVIDIA H100	
Réglage fin du modèle			
Inférence			

Tâche	Serveur/Instance	Processeurs graphiques par serveur/instance	Autres achats
Solution AWS SageMaker			
Gestion des clusters	Sans objet	Sans objet	7 To de stockage EBS par mois pour les instances ml.r5.16xlarge ; 1 To d'entrée et 15 To de sortie S3
ordinateurs portables	20x ml.t3.medium	Sans objet	
Traitement des données	2 x ml.r5.16xlarge	Sans objet	
Réglage fin du modèle	**ml.p5.48xlarge	8x NVIDIA H100	
Inférence	**ml.p5.48xlarge	8x NVIDIA H100	
Solution Azure Machine Learning			
Gestion des clusters	Sans objet	Sans objet	10 000 000 opérations de transfert de données Azure Block Blob Storage
ordinateurs portables	20x D2 v2	Sans objet	
Traitement des données	**M64	s.o.	
Réglage fin du modèle	1x ND96isr H100 v5	8x NVIDIA H100	
Inférence	1x ND96isr H100 v5	8x NVIDIA H100	

Notez que cette étude utilise les tarifs des processeurs graphiques NVIDIA H100. Bien les serveurs PowerEdge XE9680 prennent en charge les processeurs graphiques H200 lancés par NVIDIA, nous avons choisi de comparer le TCO pour des solutions de configuration similaire utilisant des processeurs graphiques H100. Pour connaître les caractéristiques exactes des solutions que nous avons comparées, consultez les données scientifiques qui sous-tendent le rapport.

Pour cette analyse, nous avons essayé de créer un exemple de scénario largement applicable pour estimer les différences de coûts entre les environnements. Nous avons choisi le modèle d'IA générative Llama 3 8B, car il s'agit d'un modèle open source largement disponible. Nous avons inclus les coûts des ordinateurs portables consacrés au développement de l'apprentissage automatique par les scientifiques des données, des tâches de traitement des données, du réglage fin continu des modèles et de l'inférence en temps réel. Nous n'avons pas inclus les coûts de stockage au-delà de ceux dont les serveurs ou instances avaient besoin pour effectuer leurs tâches.

Pour les solutions Dell sur site, nous avons supposé que les ordinateurs portables de développement et les tâches de gestion du cluster s'effectueraient sur le cluster Dell PowerEdge R660, tandis que les tâches de traitement, de réglage fin et d'inférence s'effectueraient sur le cluster Dell PowerEdge XE9680.

Pour les solutions Cloud, nous avons choisi des instances adaptées aux besoins d'une tâche. Les instances d'ordinateurs portables étaient très petites, tandis que nous avons donné aux instances de traitement une mémoire importante. Étant donné que les services de Cloud public créent une nouvelle instance pour chaque tâche, chacune de ces tâches dispose d'une instance dédiée à huit processeurs graphiques pour sa durée d'exécution. Par conséquent, nous avons calculé le nombre de tâches que les serveurs PowerEdge XE9680 pouvaient effectuer tout en conservant le même rapport processeur graphique par tâche. Nous avons également ajouté une estimation des coûts de transfert de données vers et depuis le stockage en mode objet du fournisseur de Cloud pour tenir compte du coût de déplacement des données dans le Cloud.

Pour tenir compte des différentes réalités commerciales et effectuer une comparaison équitable, nous avons avancé les hypothèses suivantes :

- Les coûts ne comprennent pas les taxes, car les tarifs spécifiques varient en fonction de l'emplacement géographique.
- Tous les logiciels sont open source, avec des licences permettant une utilisation commerciale.
- Nous excluons les coûts de gestion des solutions Cloud. Pour les solutions sur site, nous prenons en compte les coûts d'administration système continus pour assurer la maintenance du matériel et le support des scientifiques des données.
- Pour les solutions sur site, nous prenons en compte les coûts liés à l'espace, à l'alimentation et au refroidissement du datacenter physique.
- Pour l'achat CAPEX de Dell AI Factory, nous avons exclu tout calcul du coût d'exploitation du capital/de l'amortissement.

Pour plus de détails sur les hypothèses et les calculs, voir les données scientifiques qui ont servi à établir ce rapport.

Comparaison des coûts de l'IA générative : solutions Dell sur site Dell AI Factory par rapport au Cloud

Hypothèses pour comparer les coûts de l'IA générative

- Nous supposons qu'il y a 22 jours de travail par mois, avec des charges applicatives configurées pour s'exécuter pendant 24 heures afin d'optimiser l'utilisation.
- Par conséquent, chaque serveur offre 528 heures d'exécution par mois.
- Les tâches de traitement des données peuvent s'exécuter pendant 528 heures x deux serveurs Dell PowerEdge XE9680 = 1 056 heures d'exécution.
- Vingt analystes de données travaillent 8 heures par jour pendant 22 jours par mois, pour un total de 3 520 heures.

Puisque les tâches de traitement utilisent le processeur et la mémoire, nous les hébergeons pour les 1 056 heures d'exploitation complète des serveurs PowerEdge XE9680. Nous avons divisé les tâches de finition du modèle et d'inférence entre les deux serveurs, en supposant que la charge applicative nécessiterait plus de temps pour la finition du modèle que pour l'inférence. Par conséquent, nous avons calculé 792 heures par mois consacrées aux tâches de réglage fin et 264 heures par mois aux tâches d'inférence.

Pour finir, pour l'utilisation d'ordinateurs portables par 20 scientifiques des données, nous avons supposé que chacun avait une journée de travail standard de 8 heures pendant 5 jours par semaine, soit un total de 3 520 heures par mois. Le nombre de scientifiques des données que votre société emploie pour maintenir et affiner votre modèle dépend de plusieurs facteurs, tels que les différentes manières dont vous souhaitez interpréter votre jeu de données ou le nombre d'applications alimentées par ce dernier. Nous avons choisi un chiffre situé dans la partie supérieure de l'échelle pour représenter un coût de mise à niveau qui s'appliquerait à de nombreuses entreprises. Étant donné que ces instances dans le Cloud public sont très petites et très peu coûteuses par rapport à la solution dans son ensemble, le nombre de scientifiques des données n'aura pas d'impact important sur le coût total de notre solution. À l'aide de ces calculs de temps d'activité, nous avons pu indiquer le nombre d'heures d'exécution de chaque type d'instance par mois sur les deux solutions Cloud. Pour connaître les coûts totaux finaux de toutes les solutions, voir les données scientifiques qui ont permis d'élaborer le rapport.

Tarifs détaillés de la solution sur site Dell AI Factory

Dell a fourni une proposition commerciale basée sur le prix recommandé par Dell pour la solution sur site Dell AI Factory. Cette proposition commerciale incluait le coût des serveurs et des commutateurs, ProDeploy Plus pour les services d'installation sur site des serveurs et un plan ProSupport for Infrastructure de 5 ans fournissant les services de support et de maintenance de l'équipement. Remarque : Nous avons opté pour un plan de support de 5 ans, car bien que nous ayons limité notre coût total de propriété à 4 ans, la plupart des serveurs durent de 3 à 5 ans et ont besoin d'un service au-delà des 4 ans que nous avons envisagés. Nous avons ensuite calculé les coûts d'énergie pour l'alimentation et le refroidissement, ainsi que les coûts d'espace rack du datacenter pour une période de 4 ans, ainsi que les coûts administratifs liés à la maintenance de l'équipement pendant 4 ans.

Tarifs détaillés de la solution Cloud AWS SageMaker

AWS divise son service SageMaker en plusieurs sous-services couvrant des tâches telles que le traitement et la formation ainsi que les ordinateurs portables des scientifiques de données. Notez que, alors que nous affinons un modèle pré-entraîné, le sous-service AWS SageMaker est appelé SageMaker Training. Pour obtenir les tarifs de SageMaker, nous avons utilisé l'AWS Pricing Calculator et le calculateur Machine Learning Savings Plans.[2,3] Pour notre coût total de propriété, nous avons évalué les instances pour les ordinateurs portables, le traitement, le réglage fin du modèle et l'inférence comme suit :

Tableau 2 : Instances d'environnement AWS SageMaker et heures d'exécution par mois.

Modèle d'instance	Nombre d'instances	Tâche	Temps d'exécution (heures/mois)/ instance
**ml.t3.medium	20	Ordinateur portable de scientifique des données	176
ml.r5.16xlarge	2	Traitement des données	1 056
ml.p5.48xlarge	1	Réglage fin du modèle	792
ml.p5.48xlarge	1	Inférence	264

Hypothèses relatives aux détails de tarification :

Nous avons choisi deux instances ml.r5.16xlarge pour le traitement des données afin de garantir au moins 1 To de mémoire par tâche, d'après des recherches indiquant que les tâches de traitement consomment beaucoup de mémoire.[4,5]

- Nous avons ajouté 3,5 To par mois de stockage EBS à chaque instance ml.r5.16xlarge, car elles ne sont pas fournies avec des disques.
- Même si nous n'avons pas estimé les coûts du stockage hébergeant le jeu de données principal, nous avons estimé les coûts de transfert de données S3 à 1 To d'entrée et 15 To de sortie par mois pour tenir compte des sous-jeux de données que les tâches d'entraînement et d'inférence utiliseront.
- Les instances ml.p5.48.large étaient équipées d'un stockage NVMe à attachement direct. Nous n'avons donc pas ajouté de stockage EBS pour ces instances.

Remarque : SageMaker inclut un adaptateur EFA (Elastic Fabric adapter) qui offre des débits élevés.[6] Bien que nous estimions que la gestion de réseau de la solution Dell est adaptée à notre scénario, vous pouvez opter pour une configuration réseau avec plus de bande passante. Par conséquent, il est possible que la solution AWS traite plus de tâches que la solution Dell en fonction de vos choix de gestion de réseau.

AWS propose des tarifs à la demande et des plans d'économies SageMaker. La tarification à la demande est la plus coûteuse, tandis que les plans d'économies offrent jusqu'à 64 % de réduction des coûts avec un engagement de 3 ans.[7] AWS ne propose pas de tarification spécifique pour un engagement de 4 ans. Par conséquent, pour travailler en fonction du meilleur coût possible pour notre TCO de 4 ans, nous avons calculé le prix de la configuration AWS au prorata sur 4 ans en utilisant le tarif avec engagement de 3 ans.[8] (Pour un autre aperçu de la tarification AWS, nous avons également calculé les coûts sur 4 ans en utilisant 3 ans au prix d'engagement sur 3 ans plus 1 an au prix d'engagement sur 1 an. Voir le rapport scientifique pour obtenir ces résultats supplémentaires.) En outre, AWS offre aux clients la possibilité de payer les coûts à l'avance pour une réduction plus importante des coûts, ce que nous avons choisi de faire pour nos calculs du coût total de propriété. Notez que nous avons basé l'évaluation du prix de notre solution AWS sur la région est des États-Unis (Ohio) et que la tarification peut varier selon la région.

Moins de dépenses avec une solution sur site Dell AI Factory par rapport à AWS SageMaker

En utilisant les hypothèses ci-dessus pour les deux solutions, nous avons calculé une comparaison du coût total de propriété sur 4 ans. Nos calculs montrent que le choix de la solution sur site Dell AI Factory pour exécuter des charges applicatives d'IA générative peut offrir de réelles économies par rapport à l'exécution de la même charge applicative sur AWS SageMaker.

Comme le montre la Figure 2, nous avons calculé que la solution sur site Dell AI Factory coûterait 71 % moins cher qu'une solution AWS SageMaker similaire. Étant donné que le coût de la solution AWS est 3,5 fois plus élevé que sur 4 ans, les utilisateurs peuvent supposer qu'ils atteindraient presque leur seuil de rentabilité à 1 an avec une solution sur site Dell AI Factory sur site par rapport au coût de l'hébergement AWS.

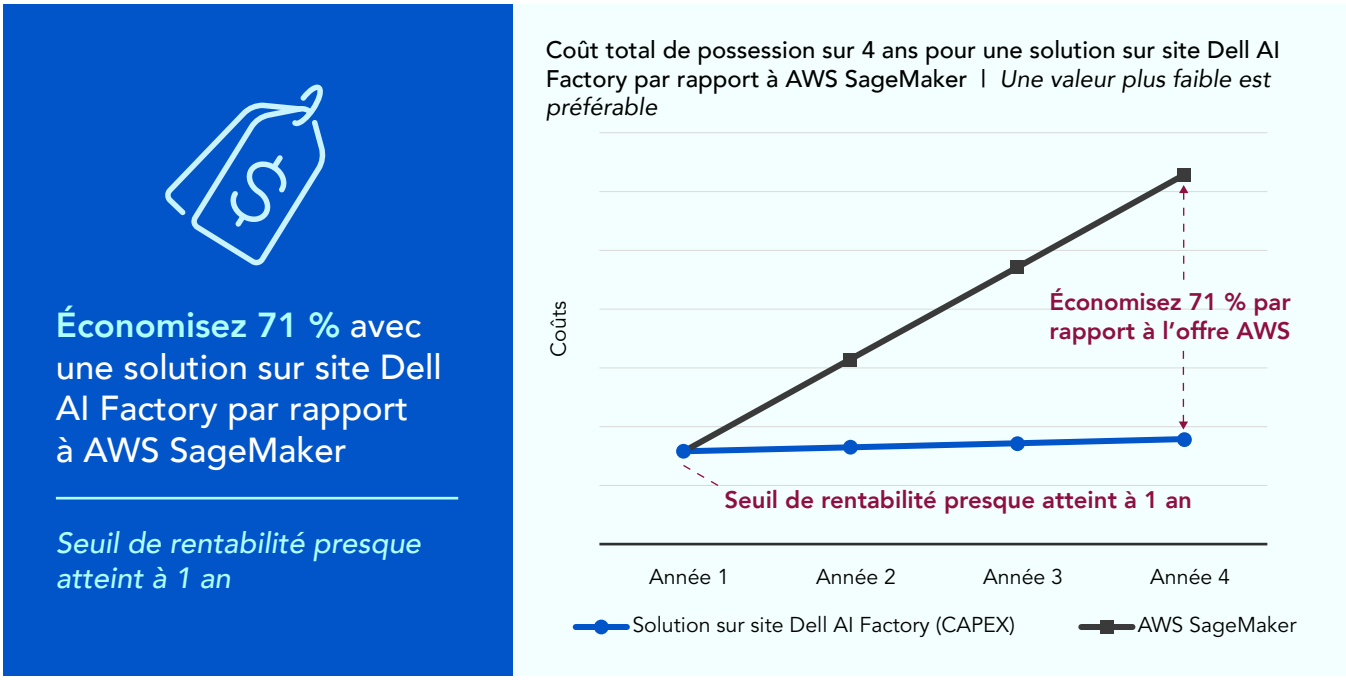


Figure 2 : Coûts relatifs d'une solution sur site Dell AI Factory et d'une solution AWS SageMaker sur 4 ans.

Tarifs de la solution Cloud Azure Machine Learning

Pour l'environnement de service Azure Machine Learning, nous avons choisi des instances pour les quatre mêmes tâches que l'environnement AWS : les ordinateurs portables de développeurs des scientifiques des données, le traitement des données, le réglage fin et l'inférence. Nous avons obtenu nos tarifs via l'Azure Pricing Calculator, en choisissant l'option du plan d'économies réservées sur 4 ans.[9] Les instances dont le tarif a été évalué sont les suivantes :

Tableau 3 : Instances de l'environnement Azure Machine Learning et temps d'exécution par mois.

Modèle d'instance	Nombre d'instances	Tâche	Temps d'exécution (heures/mois/instance)
D2 v2	20	Ordinateur portable de scientifique des données	176
M64	1	Traitement des données	1 056
ND96isr H100 v5	1	Réglage fin du modèle	792
ND96isr H100 v5	1	Inférence	264

Détails des tarifs pour les hypothèses Azure Machine Learning

- Toutes les instances Azure Machine Learning sont fournies avec un stockage en mode bloc rattaché. Nous n'avons donc pas évalué de stockage supplémentaire pour l'environnement Azure. Comme dans nos calculs pour AWS, nous avons effectué environ 10 000 000 opérations de transfert de données Block Blob Storage pour transférer des données vers et depuis les instances Machine Learning. La calculatrice pour ces transactions inclut plusieurs transactions spécifiques, telles que les opérations d'écriture, les opérations de lecture, etc. Nous avons choisi 10 000 000 pour chacune.

Azure propose des tarifs en Pay as you Go, des plans d'économies Azure et les options Azure Reservations pour le service Machine Learning.[10] À l'instar d'AWS, Azure propose une option de paiement à l'avance, mais il ne semble pas modifier le coût mensuel ni fournir de remise. Pour correspondre au mieux à la tarification de l'environnement AWS, nous avons choisi le plan Reservations de 3 ans et effectué un calcul sur 4 ans au prorata. (Comme pour AWS, nous avons également calculé les coûts et les économies pour une tarification réservée de 3 ans et une tarification réservée de 1 an, que vous pouvez voir en consultant les données scientifiques qui ont servi à établir ce rapport.) Comme pour AWS, nous avons basé l'évaluation du prix de notre solution Microsoft Azure sur la région est 2 des États-Unis. Notez que la tarification peut varier en fonction de la région.

Un seuil de rentabilité atteint en moins de deux ans avec une solution sur Dell AI Factory en remplacement d'Azure ML

À l'aide des hypothèses ci-dessus, nous avons calculé les coûts d'une solution Azure sur 4 ans et les avons comparés à nos estimations du coût total de propriété sur 4 ans pour la solution sur site Dell AI Factory. À nouveau, nos calculs montrent que la solution sur site Dell AI Factory pour les charges applicatives d'IA générative peut offrir des économies significatives sur 4 ans par rapport à une solution Azure ML comparable.

En réalité, nous estimons que la solution sur site Dell AI Factory coûte 61 % moins cher qu'une solution Azure Machine Learning similaire (voir Figure 3). Ces résultats montrent que la conservation de votre matériel en interne pour l'IA générative avec une solution sur site Dell AI Factory peut vous aider à rendre votre budget d'IA générative plus raisonnable. La solution Azure ML étant plus de 2,5 fois plus chère sur 4 ans, les clients de la solution sur site Dell AI Factory peuvent s'attendre à atteindre un seuil de rentabilité à environ 1 an et demi par rapport à la tarification Azure.

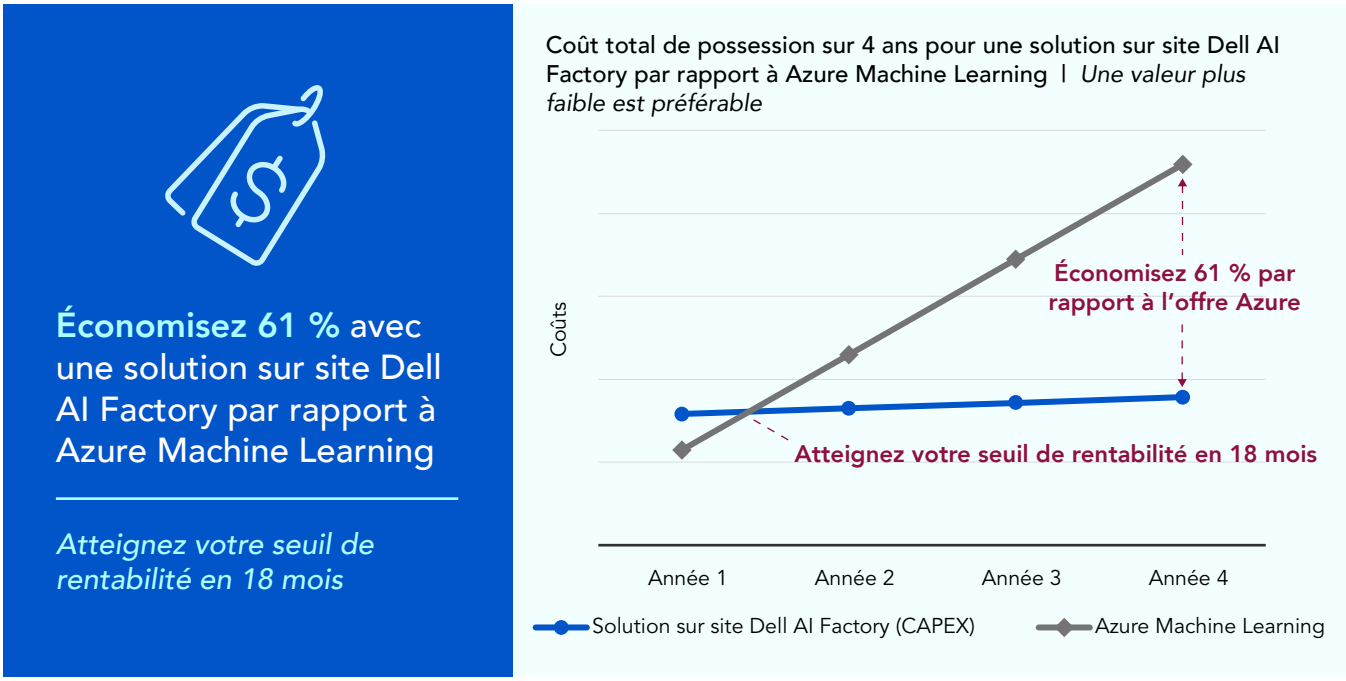


Figure 3 : Coûts relatifs d'une solution sur site Dell AI Factory et d'une solution Azure Machine Learning sur 4 ans.

Faites des économies en choisissant les abonnements Dell APEX

Certaines organisations peuvent trouver l'engagement à long terme inhérent à une solution sur site traditionnelle prohibitif. C'est à cette fin que Dell propose les abonnements Dell APEX. Dell peut installer du matériel dans le datacenter de votre entreprise afin qu'il reste sur site comme la solution traditionnelle. Il propose un engagement de 3, 4 ou 5 ans pour les ressources de calcul à un taux de consommation spécifié pour un paiement mensuel cohérent. Si vous avez besoin d'un niveau de consommation supérieur à votre engagement, vous pouvez puiser dans les ressources restantes moyennant un coût supplémentaire. À la fin de votre abonnement, vous pouvez annuler le service et renvoyer le matériel, renouveler l'abonnement en l'état ou passer à une solution qui répond le mieux à vos besoins.[11]

Pour la comparaison de notre coût total de propriété, nous avons reçu une proposition commerciale de Dell pour le matériel inclus dans notre solution sur site CAPEX Dell AI Factory, mais également pour l'ajout d'un abonnement de 4 ans à des abonnements Dell APEX à un taux de consommation garanti de 75 %. Les taux de consommation des abonnements Dell APEX pour les serveurs sont basés sur la durée pendant laquelle un serveur utilise plus de 5 % d'activité du processeur au cours d'un mois.

Suppositions pour les abonnements Dell APEX

- Environ 726 heures par mois avec un taux de consommation garanti de 75 % = maximum de 544,5 heures de temps serveur par mois avant d'avoir besoin de ressources supplémentaires. Par souci de cohérence avec les autres calculs, nous avons utilisé 528 heures par mois.
- L'estimation incluait également les plans ProDeploy Plus et ProSupport Next-Business Day, nous n'avons donc pas inclus les coûts d'administration pour la configuration initiale.
- Nous avons inclus les mêmes coûts d'alimentation, de refroidissement et d'espace rack du datacenter que pour notre solution traditionnelle.

Nous avons conclu que les abonnements Dell APEX, qui associent les avantages de sécurité et de contrôle d'une solution sur site traditionnelle à la commodité et à la flexibilité d'un service managé, pouvaient permettre aux entreprises d'économiser beaucoup sur 4 ans, par rapport aux solutions Cloud que nous avions évaluées.

Comme le montre la Figure 4, les abonnements Dell APEX coûtent 71 % moins cher que la solution AWS SageMaker. La solution AWS coûte 3,5 fois plus cher que les abonnements Dell APEX sur 4 ans. Avec les abonnements Dell APEX, vous pouvez payer moins dès le premier jour et dépenser beaucoup moins sur 4 ans.

Coût total de possession sur 4 ans pour Dell APEX Subscriptions par rapport à AWS SageMaker | *Une valeur plus faible est préférable*

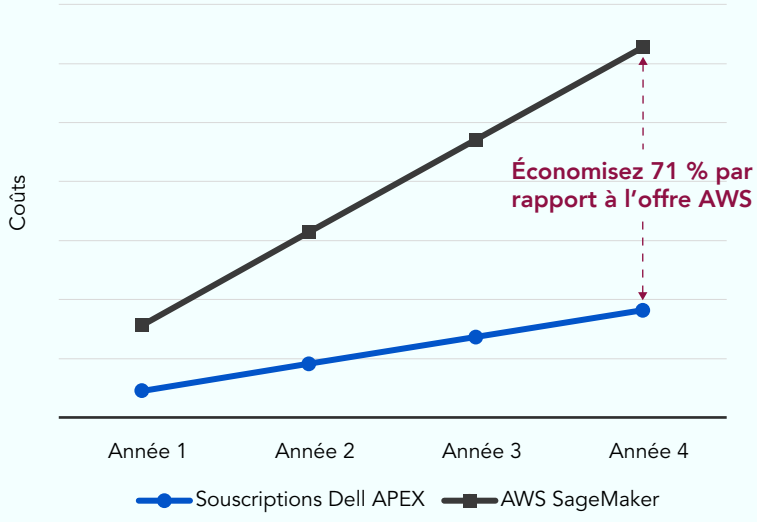
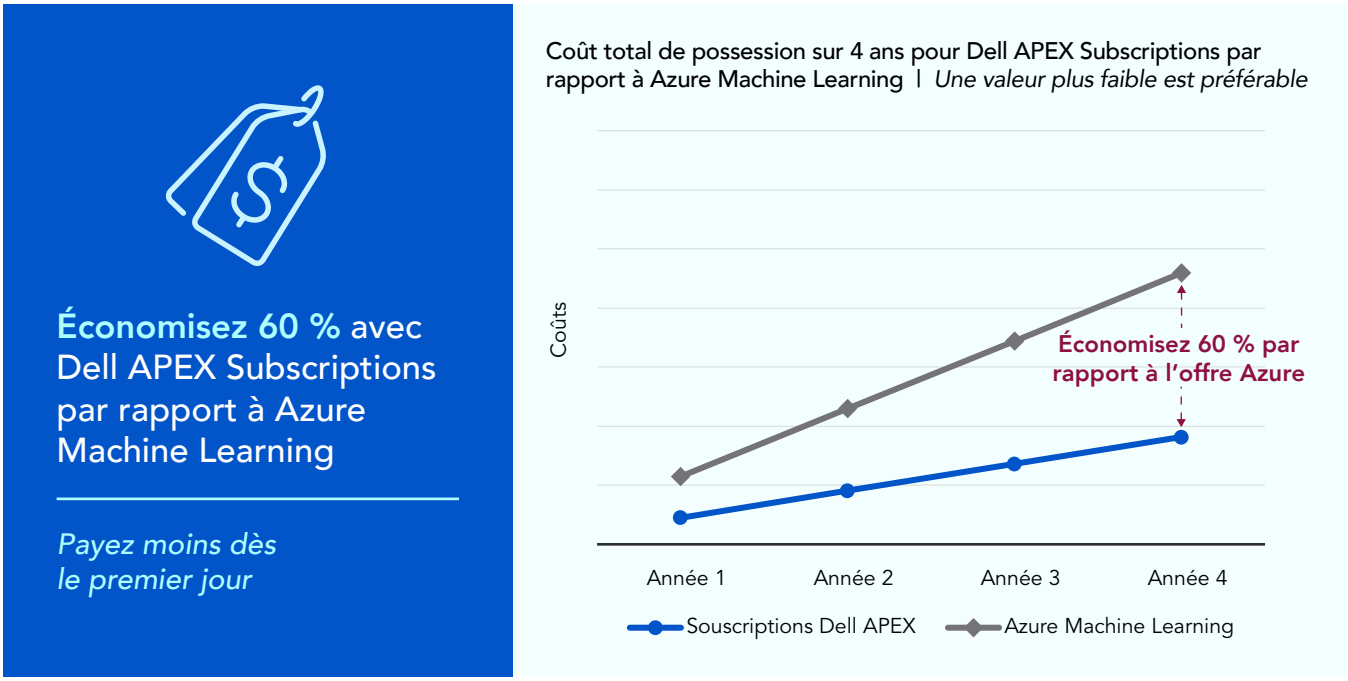


Figure 4 : Coûts relatifs d'abonnements Dell APEX et d'une solution AWS SageMaker sur 4 ans.

Comme le montre la Figure 5, les abonnements Dell APEX ont également permis de réaliser des économies significatives par rapport à la solution Azure Machine Learning, réduisant le TCO sur 4 ans de 60 %. Cela signifie que la solution Azure Machine Learning coûterait 2,5 fois plus cher que l'utilisation d'abonnements Dell APEX sur 4 ans. Ces résultats montrent que les entreprises soucieuses de leur budget et cherchant à implémenter l'IA générative peuvent tout à fait répondre à leurs attentes en utilisant des abonnements Dell APEX, plutôt que d'héberger ces charges applicatives potentiellement sensibles dans le Cloud. En outre, comme dans la



comparaison précédente, les clients paient moins dès le premier jour où ils souscrivent des abonnements Dell APEX, ce qui se traduit par une réduction considérable des coûts sur une période de 4 ans.

Figure 5 : Coûts relatifs d'abonnements Dell APEX et d'une solution Azure Machine Learning sur 4 ans.

Autres considérations pour l'exécution de charges applicatives d'IA générative sur site plutôt que dans le Cloud

Placer de grandes quantités de données utilisateur dans le Cloud public à des fins de collecte et d'affinage sur des plateformes tierces peut présenter des risques de sécurité importants, notamment :

- Exposer les données à des interfaces publiques auxquelles les pirates peuvent accéder. Par exemple, CrowdStrike a découvert l'une de ces failles de sécurité qui lui permettait de trouver des buckets AWS S3 en fonction des demandes DNS.[12]
- Une complexité accrue pouvant entraîner des erreurs de configuration, du fait des équipes IT qui doivent jongler avec plusieurs services et fournisseurs Cloud qui modifient régulièrement les configurations et paramètres par défaut.
- Erreur humaine amplifiée lors de l'utilisation d'API basées sur le Cloud susceptibles d'exposer des données sensibles.[13]

Conserver des LMS sur des réseaux privés peut atténuer ces risques, car les solutions internes disposent d'un meilleur contrôle sur les flux de données, l'isolation du réseau, les contrôles des API, le maintien de la conformité des données, l'optimisation des performances, etc. En outre, les utilisateurs exécutant des LLM localement ont plus de contrôle sur l'ensemble de la pile, du matériel sur lequel s'exécute le LLM au modèle et aux données qui activent la solution. Les administrateurs peuvent suivre une formation supplémentaire pour s'assurer que les LLM

locaux sont conformes aux réglementations spécifiques. Dans le Cloud, les utilisateurs ont moins de contrôle sur l'infrastructure et l'implémentation sous-jacentes.[14] En outre, les solutions sur site permettent de maintenir les coûts prévisibles au lieu de varier d'un mois à l'autre.

Le stockage et le transfert de données constituent une grande partie des exigences des applications LLM. L'entraînement d'un LLM nécessite de grandes quantités de données, qui doivent être stockées quelque part, puis être transférées entre le stockage et les ressources de calcul à des fins de traitement. Si les appareils, les bases de données et les données utilisateur alimentant votre LLM stockent déjà leurs données sur site, les coûts de transition de ces données vers le Cloud et la bande passante réseau nécessaire peuvent être élevés.

Modèles Llama 3

Llama 3, qui signifie Large Language Model Meta AI, est une technologie de traitement du langage libre et polyvalente développée par Meta. Il s'agit d'un grand modèle de langage (LLM) pré-entraîné avec deux variantes de taille de modèle principales basées sur le nombre de paramètres (8B et 70B), adaptées à de nombreux cas d'utilisation.[15] Llama 3 a été entraîné par Meta avec un « ... nouvel ensemble d'évaluation humaine de haute qualité. Cet ensemble d'évaluation contient 1 800 invites qui couvrent 12 cas d'utilisation clés : demander des conseils, brainstorming, classification, répondre à des questions fermées, codage, écriture créative, extraction, habiter un personnage/profil, répondre à des questions ouvertes, raisonner, réécrire et résumer. »[16]

Découvrez-en plus sur Llama 3 en accédant à https://ai.meta.com/blog/meta-llama-3/.

Conclusion

D'après notre étude, l'hébergement de charges applicatives d'IA générative sur site, dans une solution Dell traditionnelle ou à l'aide d'abonnements Dell APEX, pourrait considérablement réduire vos coûts d'IA générative sur 4 ans par rapport à l'hébergement de ces charges applicatives dans le Cloud. En fait, nous avons constaté que la solution sur site Dell AI Factory pourrait réduire les coûts de 71 % par rapport à une solution AWS SageMaker comparable et de 61 % par rapport à une solution Azure ML comparable. Ces résultats montrent que les entreprises qui cherchent à implémenter l'IA générative et à en tirer parti peuvent trouver de nombreux avantages dans une solution sur site Dell AI Factory, qu'elles choisissent de l'acheter et de la gérer elles-mêmes ou qu'elles optent pour des abonnements Dell APEX. Choisir une solution sur site Dell AI Factory peut permettre à votre entreprise de réaliser d'importantes économies par rapport à un hébergement de l'IA générative dans le Cloud : vous contrôlez en effet la sécurité et la confidentialité de vos données, des mises à jour et des modifications apportées à l'environnement, tout en assurant une gestion cohérente de votre environnement.

---

1. CIO, « How to get gen AI spend under control », consulté le 7 avril 2025, https://www.cio.com/article/3478467/how-to-get-gen-ai-spend-under-control.html.
2. AWS, « AWS Pricing Calculator », consulté le 16 avril 2025, https://calculator.aws/#/.
3. AWS, « Machine Learning Savings Plans », consulté le 16 avril 2025, https://aws.amazon.com/savingsplans/ml-pricing/.
4. StackOverflow, « Why should preprocessing be done on CPU rather than GPU? », consulté le 16 avril 2025, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu.
5. Hugging Face, « Model Memory Requirements », consulté le 16 avril 2025, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
6. AWS, « Training large language models on Amazon SageMaker: Best practices », consulté le 16 avril 2025, https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/.

7. AWS, « Machine Learning Savings Plans », consulté le 16 avril 2025, https://aws.amazon.com/savingsplans/ml-pricing/.
8. Remarque : AWS a confirmé que l'instance ml.p5.48xlarge est incluse dans l'abonnement d'engagement de 3 ans. Au moment de cette étude, il n'était pas répertorié dans le calculateur de plan d'économies. Nous avons estimé le coût de l'instance ml.p5 en utilisant le pourcentage d'économies répertorié pour la version p5.48xlarge sans apprentissage automatique, comme indiqué sur https://aws.amazon.com/savingsplans/compute-pricing/.
9. Microsoft, « Azure Pricing Calculator », consulté le 16 avril 2025, https://azure.microsoft.com/en-us/pricing/calculator/.
10. Microsoft, « Azure Machine Learning Pricing », consulté le 16 avril 2025, https://azure.microsoft.com/en-us/pricing/details/machine-learning/.
11. Dell, « Dell APEX Subscriptions », consulté le 16 avril 2025, https://www.dell.com/en-us/dt/apex/subscriptions.htm.
12. CrowdStrike, « 12 Cloud Security Issues: Risks, Threats, and Challenges », consulté le 16 avril 2025, https://www.crowdstrike.com/cybersecurity-101/cloud-security/cloud-security-risks-threats-challenges/.
13. CrowdStrike, « 12 Cloud Security Issues: Risks, Threats, and Challenges ».
14. DataCamp, « Avantages et inconvénients de l'utilisation des LLM dans le Cloud par rapport à l'exécution locale des LLM », consulté le 16 avril 2025, https://www.datacamp.com/fr/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally.
15. Meta, « Introducing Meta Llama 3: The most capable openly available LMM to date », consulté le 16 avril 2025, https://ai.meta.com/blog/meta-llama-3/.
16. Meta, « Introducing Meta Llama 3: The most capable openly available LMM to date. »

Consultez les données scientifiques qui sous-tendent ce rapport ▶

▶ Consultez la version d'origine (en anglais) de ce rapport

Principled Technologies®

Facts matter.®

Ce projet a été réalisé à la demande de Dell Technologies.