Investir dans l'IA générative : analyse coûts-avantages des déploiements Dell sur site par rapport aux déploiements AWS et Azure similaires

Dans le monde technologique, l'IA générative (GenAI) ouvre de nouveaux horizons. À mesure que les entreprises du monde entier commencent à explorer comment l'IA générative peut contribuer à atteindre leurs objectifs commerciaux, elles rencontrent de nombreux obstacles pour mettre en œuvre une solution d'IA générative capable de répondre à leurs besoins spécifiques. Parmi les plus grands défis, il y a l'estimation précise du coût d'une solution GenAI de taille appropriée et le fait de savoir si elle doit être déployée sur site ou dans le Cloud. Selon Frances Karamouzis, analyste chez Gartner, « l'une des plus grandes menaces pesant sur la réussite de l'IA et de l'IA générative est le coût. Plus de la moitié des organisations abandonnent leurs efforts en raison d'erreurs dans l'estimation et le calcul des coûts ».[1]

Pour fournir aux organisations un point de départ pour 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 3 ans de deux solutions Dell™ sur site exploitant le matériel PowerEdge™ R660 et PowerEdge XE9680 (une solution traditionnelle 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, la solution Dell APEX Pay-Per-Use était la plus rentable des solutions sur 3 ans, parmi celles que nous avons comparées. Les solutions Cloud concurrentes d'AWS et Azure coûtent jusqu'à 3,81 fois plus cher que la solution Dell APEX Pay-Per-Use basée sur abonnement. Par rapport à la solution Dell traditionnelle sur site, les solutions Cloud AWS et Azure que nous avons évaluées coûteraient jusqu'à 2,88 fois plus cher. Lisez la suite pour découvrir comment l'IA générative peut aider votre entreprise et comment nous avons calculé nos résultats de coût total de possession (TCO).

Payez moins pour l'IA générative avec une solution Dell APEX de paiement à l'utilisation et une solution Dell d'IA générative sur site

Les solutions Cloud AWS et Azure concurrentes coûtent plus cher :

Plus de 3,81 fois le coût d'une solution Dell APEX en paiement à l'utilisation

Plus de 2,88 fois plus cher que la solution Dell traditionnelle sur site

Avantages de l'utilisation de modèles préentraînés pour l'IA générative

Les modèles d'intelligence artificielle (IA) sont des systèmes qui visent à imiter certains aspects de l'intelligence ou du comportement humain. Les entreprises et les particuliers utilisent des outils d'IA générative (tels que ChatGPT et DALL-E) pour générer du contenu, y compris du texte, des fichiers audio, des vidéos, des images, du code et des simulations, ainsi que des sorties plus complexes telles que du contenu marketing personnalisé, des applications personnalisées et des logiciels.[2] Une entreprise qui cherche à intégrer l'IA générative dans ses opérations peut choisir entre utiliser un modèle préentraîné ou créer son propre modèle à partir de zéro. Un modèle pré-entraîné, tel que Llama 2, est déjà entraîné sur un jeu de données de base pour l'utilisation prévue du modèle. À partir de là, les entreprises peuvent affiner le modèle à l'aide de leurs données spécifiques, ce qui permet de lancer le processus de création d'un modèle personnalisé en fonction de leurs données et de leurs cas d'utilisation[3] . Selon une source, l'utilisation de modèles préentraînés pourrait réduire le temps nécessaire à la mise en place d'un modèle d'IA fonctionnel en faisant gagner jusqu'à un an, tout en économisant des centaines de milliers de dollars[4].

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 2 13B 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 scientifique des données, 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 29 mars 2024. 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 traditionnelle sur site			
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 PowerEdge XE9680	8x NVIDIA H100	
Réglage fin du modèle			
Inférence			
Solution Dell APEX Pay-Per-Use gérée sur site			
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 PowerEdge XE9680	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	4x ND96amsr A100 v4	8x NVIDIA A100	
Inférence	4x ND96amsr A100 v4	8x NVIDIA A100	

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 2 13B, 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.
- 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 du datacenter physique, à l'alimentation et au refroidissement. Nous les intégrons en coûts d'instance pour les solutions Cloud.

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 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 la nuit 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 scientifiques des 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 développeurs, 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 Dell traditionnelle sur site

Nous avons contacté Dell et demandé une proposition commerciale recommandée pour notre solution sur site traditionnelle. 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 possession à 3 ans, la plupart des serveurs durent de 3 à 5 ans et ont besoin d'un service au-delà des trois 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 3 ans, ainsi que les coûts administratifs liés à la maintenance de l'équipement pendant 3 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.[56] 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[7, 8].
- Nous avons ajouté 7 To par mois de stockage EBS aux instances 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.[9] 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.[10] Nous avons calculé le prix de la configuration AWS en utilisant le tarif avec engagement de 3 ans.[11] 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.

Comparaison de la solution Dell traditionnelle sur site à AWS SageMaker

En utilisant les hypothèses ci-dessus pour les deux solutions, nous avons calculé une comparaison du coût total de propriété sur 3 ans. Nos calculs montrent que le choix de la solution Dell PowerEdge traditionnelle sur site 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 1, nous avons calculé que la solution AWS SageMaker pouvait coûter jusqu'à 2,88 fois plus cher que la solution Dell sur site. Remarque : Pour plus d'informations sur ce coût et sur tous les détails suivants, consultez les données scientifiques qui ont permis d'établir le rapport.

AWS SageMaker pourrait coûter presque 3 fois plus cher qu'une solution Dell similaire

Seuil de rentabilité presque atteint à 1 an

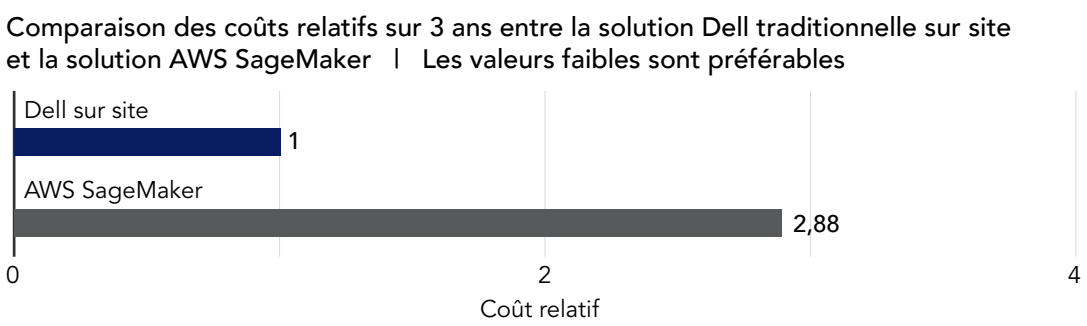

Figure 1 : Coûts relatifs d'une solution Dell d'IA générative sur site et d'une solution AWS SageMaker sur 3 ans.

Compte tenu du coût 2,8 fois plus cher sur 3 ans, les utilisateurs peuvent supposer que, même avec les coûts d'hébergement, de refroidissement et de gestion d'une solution sur site, ils atteindraient presque le seuil de rentabilité à 1 an par rapport au coût de l'hébergement AWS.

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 3 ans.[12] 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
ND96asmr A100 v4	4	Réglage fin du modèle	792
ND96asmr A100 v4	4	Inférence	264

Détails des tarifs pour les hypothèses Azure Machine Learning

- Le service Azure machine Learning n'offrait pas d'instance avec les processeurs graphiques NVIDIA H100. Nous avons donc choisi quatre instances de processeur graphique A100 pour obtenir des performances similaires.[13]
- 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.

Azure propose des tarifs en Pay as you Go, des plans d'économies Azure et les options Azure Reservations pour le service Machine Learning.[14] Azure n'offrait pas d'option de paiement initial moins chère, contrairement à AWS. Pour correspondre au mieux à la tarification de l'environnement AWS, nous avons choisi le plan Reservations de 3 ans.

Comparaison entre la solution Dell sur site et Azure ML

À l'aide des hypothèses ci-dessus, nous avons calculé les coûts d'une solution Azure sur 3 ans et les avons comparés à nos estimations du coût total de propriété sur 3 ans pour la solution Dell sur site. À nouveau, nos calculs montrent que la solution Dell PowerEdge traditionnelle sur site pour les charges applicatives d'IA générative peut offrir des économies significatives sur 3 ans par rapport à une solution Azure ML comparable.

En fait, nous estimons que le coût total de la solution Azure Machine Learning sur 3 ans serait 2,72 fois plus élevé que celui de la solution Dell traditionnelle sur site (voir Figure 2). Ces résultats montrent que conserver votre matériel en interne pour l'IA générative avec une solution Dell traditionnelle peut vous aider à rendre votre budget d'IA générative raisonnable, vous permettant ainsi d'utiliser ces économies pour innover ailleurs.

Azure Machine Learning pourrait coûter presque 3 fois plus cher qu'une solution Dell similaire

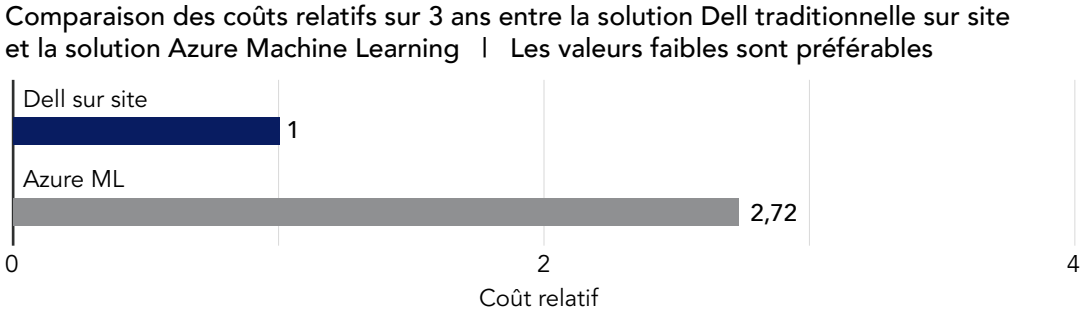

Figure 2 : Coûts relatifs d'une solution Dell sur site d'IA générative et d'une solution Azure Machine Learning sur 3 ans.

Tout comme avec la solution AWS, plus de 2,7 fois plus chère sur 3 ans, les clients Dell sur site peuvent s'attendre à atteindre un seuil de rentabilité à 1 an par rapport à la tarification Azure.

Plus d'économies en souscrivant à une solution Dell APEX Pay-Per-Use

Certaines organisations peuvent trouver l'engagement à long terme inhérent à une solution sur site traditionnelle prohibitif. C'est pourquoi Dell propose une solution Dell APEX Pay-per-use. Dell peut installer du matériel dans le datacenter de votre organisation, afin qu'il reste sur site comme la solution traditionnelle. Il propose un engagement de 3, 4 ou 5 ans via une solution Dell APEX Pay-Per-Use 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.[15]

Pour la comparaison de notre coût total de possession, nous avons reçu une proposition commerciale de Dell pour le matériel inclus dans notre environnement sur site traditionnel, mais également pour l'ajout d'un abonnement de 3 ans à une solution Dell APEX Pay-Per-Use à un taux de consommation garanti de 75 %. Les taux de consommation de la solution Dell APEX Pay-Per-Use 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. Nos hypothèses étaient les suivantes :

Hypothèses relatives à la solution Dell APEX Pay-Per-Use

- 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 la solution Dell APEX Pay-Per-Use, qui associe les avantages de sécurité et de contrôle d'une solution sur site traditionnelle à la commodité et à la flexibilité d'un service managé, pouvait permettre aux organisations d'économiser beaucoup sur 3 ans, par rapport aux solutions Cloud que nous avions évaluées.

Comme le montre la Figure 3, la solution AWS SageMaker peut coûter 3,81 fois plus cher que la solution Dell APEX Pay-per-use.

AWS SageMaker pourrait coûter presque 4 fois plus cher qu'une solution Dell APEX Pay-Per-Use similaire

Atteignez le seuil de rentabilité en moins d'1 an

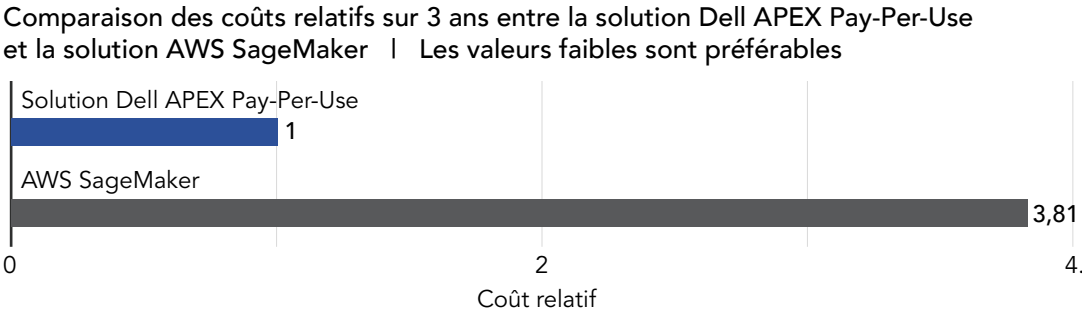

Figure 3 : Coûts relatifs d'une solution Dell APEX pay-per-use basée sur l'IA générative et d'une solution AWS SageMaker sur une période de 3 ans.

Compte tenu du prix 3,8 fois plus élevé qu'une solution Dell APEX Pay-Per-Use sur 3 ans, les utilisateurs peuvent s'attendre à atteindre le seuil de rentabilité avant la fin de la première année. Les économies d'une solution Dell APEX Pay-Per-Use étaient similaires à celles déterminées dans la comparaison à la solution Azure Machine Learning, qui coûtait 3,60 fois plus cher que la solution Dell APEX Pay-Per-Use (voir Figure 4).

Azure Machine Learning pourrait coûter presque 4 fois plus cher qu'une solution Dell APEX Pay-Per-Use similaire

Atteignez le seuil de rentabilité en moins d'1 an

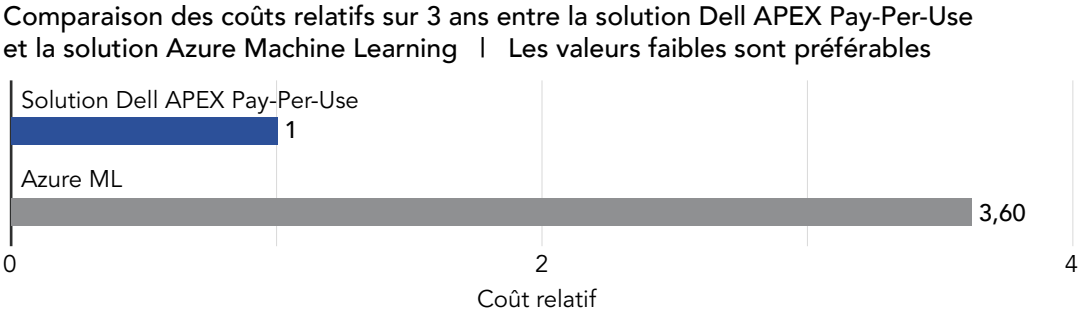

Figure 4 : Coûts relatifs d'une solution Dell APEX pay-per-use basée sur l'IA générative et d'une solution Azure Machine Learning sur une période de 3 ans.

Ces résultats montrent que les organisations soucieuses de leur budget et cherchant à implémenter l'IA générative peuvent tout à fait répondre à leurs attentes en choisissant une solution Dell APEX Pay-per-Use sur site, plutôt que d'héberger ces charges applicatives potentiellement sensibles dans le Cloud. En outre, comme pour la comparaison avec AWS, les clients peuvent également s'attendre à atteindre le seuil de rentabilité avant la fin de la première année comparé à la solution Azure.

Considérations supplémentaires pour les charges applicatives d'IA générative

Autres avantages du choix d'un système sur site plutôt que d'un système basé sur le Cloud.

Bien que l'hébergement Cloud offre des avantages tels que l'évolutivité et la flexibilité, il existe plusieurs préoccupations à prendre en compte en plus du coût, avant d'héberger vos grands modèles de langage (LLM) sur un Cloud public. L'un des plus importants est le risque associé à l'hébergement de grandes quantités de données utilisateur sur une plate-forme tierce. De nombreux LLM collectent des données utilisateur pour améliorer leurs modèles, et ces données doivent se trouver quelque part pour que les modèles puissent y accéder. Stocker ces données sensibles dans le Cloud peut les exposer à des risques tels que :

- 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.[16]
- 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.[17]

Les LLM privés réduisent ces risques, car les utilisateurs disposent généralement d'un meilleur contrôle sur leurs datacenters et donc sur leurs flux de données, l'isolation du réseau, les contrôles API, etc. En outre, les utilisateurs qui exécutent des LLM localement ont plus de contrôle sur l'ensemble de la pile, depuis le matériel sur lequel le LLM s'exécute jusqu'au modèle et aux données qui permettent d'activer 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.[18] 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 et ensuite transférées depuis le stockage vers les ressources de calcul pour le 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.

Gamme Dell AI

Il est évident que la mise en œuvre de charges applicatives d'IA ne se limite pas à la simple détermination des coûts globaux de la solution. Vous devez déterminer le modèle le mieux adapté à vos besoins et concevoir une solution capable de gérer la quantité de données dont vous disposez. Après avoir sélectionné le bon modèle et décidé d'une solution, il faut gérer les problèmes de personnel. Vous devez former ou embaucher du personnel pour vous assurer que votre personnel IT comprend la science des données, afin d'entraîner et exécuter correctement votre solution d'IA.

Dell propose une gamme complète d'IA qui ne se contente pas de vous permettre de réaliser des économies sur le matériel : elle fournit les outils dont vous avez besoin pour relever les défis lors de la mise en œuvre de l'IA. La gamme d'IA Dell couvre le matériel, les services de gestion des données, les services professionnels, la formation à l'IA, les architectures de référence et de nombreux partenariats avec d'autres entreprises axées sur l'IA.[19] Dell peut vous aider à concevoir, planifier et mettre en œuvre une solution d'IA adaptée à vos besoins. Pour en savoir plus sur la façon dont la gamme d'IA Dell se compare aux offres concurrentes, vous pouvez lire nos rapports comparant la gamme d'IA Dell à celles de Supermicro[20] et HPE.[21]

Modèles Llama 2

Llama 2, 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 trois variantes de taille de modèle principales basées sur le nombre de paramètres (7B, 13B et 70B), chacun fournissant des caractéristiques de performances différentes.[22] Llama 2 a été entraîné par Meta avec une approche d'apprentissage renforcée, en vue de produire des résultats adaptés aux familles, dans le but de se familiariser avec les choix et les préférences des humains.[23] Découvrez-en plus sur Llama 2 en accédant à https://llama.meta.com/llama2/.

Conclusion

Plonger dans le monde de l'IA générative peut potentiellement générer de nombreux avantages pour votre organisation, mais vous devez d'abord réfléchir à la meilleure façon d'implémenter ces charges applicatives d'IA générative. Que vos objectifs en matière d'IA soient de créer un chatbot pour les visiteurs en ligne, de générer des ressources marketing, d'aider au dépannage ou bien plus encore, la mise en œuvre d'une solution d'IA nécessite une planification et une prise de décision minutieuses. Une décision majeure est de savoir si vous souhaitez héberger l'IA générative dans le Cloud ou conserver vos données sur site. Les solutions sur site traditionnelles peuvent offrir une sécurité et un contrôle supérieurs, ce qui est une préoccupation majeure lorsque de grandes quantités de données potentiellement sensibles sont traitées. Mais la prise en charge d'une solution d'IA générative sur site pèsera-t-elle sur le budget IT d'une organisation ?

Dans notre étude, nous avons constaté que la proposition de valeur indique exactement le contraire : l'hébergement de charges applicatives d'IA générative sur site, dans une solution Dell traditionnelle ou à l'aide d'une solution gérée Dell APEX Pay-Per-Use, pourrait considérablement réduire vos coûts d'IA générative sur 3 ans par rapport à l'hébergement de ces charges applicatives dans le Cloud. En fait, nous avons constaté que la solution AWS SageMaker comparable coûterait jusqu'à 3,8 fois plus cher et que la solution Azure ML coûterait jusqu'à 3,6 fois plus cher que l'IA générative sur une solution Dell APEX Pay-per-Use. Ces résultats montrent que les organisations qui cherchent à implémenter l'IA générative et à en tirer parti peuvent trouver de nombreux avantages dans une solution Dell sur site, qu'elles choisissent de l'acheter et de gérer elles-mêmes ou qu'elles optent pour une solution Dell APEX Pay-per-Use basée sur un abonnement. Choisir une solution Dell sur site peut permettre à votre organisation 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. LinkedIn, post de Jacquecine Burrel, consulté le 19 avril 2024, https://www.linkedin.com/posts/jacquecine-burrell-88094714_meet-frances-karamouzis-at-gartner-data-activity-7173514796506554368-9Sie.
2. Deloitte, « The State of Generative AI in the Enterprise », consulté le 3 avril 2024, https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-generative-ai-in-enterprise.html.
3. OneAI, « The Future is Pre-trained: The Shortcut to AI Mastery », consulté le 3 avril 2024, https://oneai.com/learn/pre-trained-ai-model.
4. NVIDIA, « What Is a Pretrained AI Model? », consulté le 3 avril 2024, https://blogs.nvidia.com/blog/what-is-a-pretrained-ai-model/.

5. AWS, « AWS Pricing Calculator », consulté le 3 avril 2024, https://calculator.aws/#/.
6. AWS, « Machine Learning Savings Plans », consulté le 3 avril 2024, https://aws.amazon.com/savingsplans/ml-pricing/.
7. StackOverflow, « Why should preprocessing be done on CPU rather than GPU? », consulté le 3 avril 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu.
8. Hugging Face, « Model Memory Requirements », consulté le 3 avril 2024, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
9. AWS, « Training large language models on Amazon SageMaker: Best practices », consulté le 3 avril 2024, https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/.
10. AWS, « Machine Learning Savings Plans », consulté le 3 avril 2024, https://aws.amazon.com/savingsplans/ml-pricing/.
11. 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/.
12. Microsoft, « Azure Pricing Calculator », consulté le 3 avril 2024, https://azure.microsoft.com/en-us/pricing/calculator/.
13. Comet, « Comparison of NVIDIA A100, H100 + H200 GPUs », consulté le 3 avril 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.
14. Microsoft, « Azure Machine Learning Pricing », consulté le 3 avril 2024, https://azure.microsoft.com/en-us/pricing/details/machine-learning/.
15. Dell, « Dell APEX Flex on Demand », consulté le 3 avril 2024, https://www.delltechnologies.com/partner/en-us/partner/apex-flex-on-demand.htm.
16. CrowdStrike, « 12 Cloud Security Issues: Risks, Threats, and Challenges », consulté le 3 avril 2024, https://www.crowdstrike.com/cybersecurity-101/cloud-security/cloud-security-risks-threats-challenges/.
17. CrowdStrike, « 12 Cloud Security Issues: Risks, Threats, and Challenges ».
18. DataCamp, « Avantages et inconvénients de l'utilisation des LLM dans le Cloud par rapport à l'exécution locale des LLM », consulté le 3 avril 2024, https://www.datacamp.com/fr/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally.
19. Dell, « Dell AI solutions », consulté le 29 avril 2024, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#footnote-ref1&tab0=0.
20. Principled Technologies, « Finding the path to AI success with the Dell AI portfolio », consulté le 29 avril 2024, https://facts.pt/q9p46K9.
21. Principled Technologies, « Meeting the challenges of AI workloads with the Dell AI portfolio », consulté le 29 avril 2024, https://facts.pt/zPmSx4c.
22. Meta, « Llama 2: open source, free for research and commercial use », consulté le 3 avril 2024, https://llama.meta.com/llama2/.
23. Pavan Belagatti, « Unpacking Meta's Llama 2: The Next Leap in Generative AI », consulté le 3 avril 2024, https://www.singlestore.com/blog/a-complete-beginners-guide-to-llama2/

Données scientifiques ayant permis l'élaboration de ce rapport

Dans cette section, nous répertorions l'ensemble de nos résultats et décrivons les solutions ayant fait l'objet de tests ainsi que nos méthodologies de test.

Notre recherche s'est achevée le 29 mars 2024. Les résultats de ce rapport reflètent les configurations que nous avons finalisées et acquises pour les données de tarification le 29 mars 2024 ou avant. Inévitablement, ces configurations et leurs coûts peuvent ne pas être à jour au moment de la publication de ce rapport.

Informations sur le système

Solutions Dell sur site

Les solutions Dell traditionnelles et gérées sur site incluent toutes les deux le matériel suivant :

- 3 nœuds principaux PowerEdge R660
- 2 nœuds de travail de processeur graphique PowerEdge XE9680
- 2 infrastructures réseau PowerSwitch S5232-ON
- 1 PowerSwitch N3200-ON OOB de gestion

Tableau 1 : : informations de configuration détaillées pour chaque nœud de travail de processeur graphique PowerEdge XE9680.

Informations de configuration	Nœud de travail de processeur graphique Dell PowerEdge XE9680
Nombre de nœuds dans la solution	2
Boîtier	
Boîtier	Boîtier 6U XE9680 avec 8 processeurs graphiques 8x 2.5 NVMe uniquement
Processeur	
Nombre de processeurs	2
Fournisseur et modèle	Intel® Xeon® Platinum 8468
Nombre de cœurs (par processeur)	48 cœurs et 96 threads
Processeurs graphiques	
Nombre de processeurs graphiques	1 assemblage de 8 processeurs graphiques
Fournisseur et modèle	Montage de processeurs graphiques NVIDIA® HGX H100, 8 processeurs graphiques, SXM 80 Go 700 W
Module(s) de mémoire	
Mémoire totale du système (Go)	1 024
Nombre de modules de mémoire	16
Type	RDIMM, 4 800 MT/s, double rangée
Size (GB)	64
Contrôleur de stockage	
Fournisseur et modèle	Carte contrôleur BOSS-N1 + 2 M.2 480 Go (RAID 1)
Stockage local (type A)	
Taille totale des disques dans le système (To)	44,8

Informations de configuration	Nœud de travail de processeur graphique Dell PowerEdge XE9680
Nombre de disques	7
Taille de disque (To)	6,4
Informations sur les disques (vitesse, interface, type)	Disque AG Enterprise NVMe™, utilisation mixte, U.2 Gen 4 avec support
Carte NIC	
Nombre et type de ports	2 disques 25GbE de 10 pouces
Fournisseur et modèle	Intel E810-XXV à deux ports 10/25 GbE SFP28, OCP NIC 3.0
Carte réseau	
Nombre et type de ports	2 disques 100GbE de 10 pouces
Fournisseur et modèle	Adaptateur réseau Mellanox ConnectX-6 DX à deux ports 100 GbE QSFP56, hauteur standard
Ventilateurs de refroidissement	
Nombre de ventilateurs de refroidissement	6
Fournisseur et modèle	Ventilateur très hautes performances
Blocs d'alimentation	
Nombre de blocs d'alimentation	1
Fournisseur et modèle	Alim. entièrement redondante 5+1 (ou 3+3 tolérant aux pannes), enfichage à chaud, 2 800 W MM CA HT (200-240 V CA) Titanium, connecteur C22
Puissance de chaque bloc (W)	2800
ProSupport et ProDeploy	
ProSupport (5 ans)	ProSupport avec service sur site le jour ouvré suivant
ProDeploy Plus	ProDeploy Plus, serveur Dell série XE 5U/6U
Gestion des systèmes intégrée	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage™ Enterprise Advanced Plus

Tableau 2 : Informations de configuration détaillées pour chaque nœud principal PowerEdge R660.

Informations de configuration	Nœud principal Dell PowerEdge R660
Nombre de nœuds dans la solution	3
Boîtier	
Boîtier	Boîtier 2,5" avec jusqu'à 10 disques durs (SAS/SATA), PERC11, 1CPU
Processeur	
Nombre de processeurs	1
Fournisseur et modèle	Intel Xeon Gold 6426Y
Nombre de cœurs (par processeur)	16 cœurs et 32 threads

Informations de configuration	Nœud principal Dell PowerEdge R660
Module(s) de mémoire	
Mémoire totale du système (Go)	192
Nombre de modules de mémoire	12
Type	16
Size (GB)	RDIMM, 4 800 MT/s, rangée simple
Contrôleur de stockage	
Fournisseur et modèle	Carte contrôleur BOSS-N1 + 2SED M.2 960 Go (RAID 1)
Stockage local (type A)	
Taille totale des disques dans le système (To)	5,76
Nombre de disques	6
Taille de disque (To)	960
Informations sur les disques (vitesse, interface, type)	Disque SSD 2,5 pouces vSAS 12 Gbit/s 512e lecture intensive AG SED à enfichage à chaud, 1 écriture/jour
Carte NIC	
Nombre et type de ports	2 disques 25GbE de 10 pouces
Fournisseur et modèle	NVIDIA ConnectX-6 Lx deux ports 10/25 GbE SFP28, carte NIC No Crypto,OCP 3.0
Ventilateurs de refroidissement	
Nombre de ventilateurs de refroidissement	4
Fournisseur et modèle	Ventilateur très hautes performances
Blocs d'alimentation	
Nombre de blocs d'alimentation	1
Fournisseur et modèle	Deux blocs d'alimentation redondants (1 + 1) à enfichage à chaud, 1 100 W MM (100-240 V CA) Titanium
Puissance de chaque bloc (W)	1 100
ProSupport et ProDeploy	
ProSupport (5 ans)	ProSupport avec service sur site le jour ouvré suivant
ProDeploy Plus	ProDeploy Plus, PowerEdge série R 1U/2U
Gestion des systèmes intégrée	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage Enterprise Advanced Plus

Instances de la solution AWS SageMaker

Tableau 3 : Informations de configuration détaillées pour les instances AWS.

Informations de configuration	ml.t3.medium (ordinateurs portables)	ml.r5.16xlarge (traitement)	ml.p5.48xlarge (inférence et réglage fin)
Nombre d'instances	20	2	2
Fournisseur de services Cloud (CSP)	AWS	AWS	AWS
Région	États-Unis, Est (Ohio)	États-Unis, Est (Ohio)	États-Unis, Est (Ohio)

Informations de configuration	ml.t3.medium (ordinateurs portables)	ml.r5.16xlarge (traitement)	ml.p5.48xlarge (inférence et réglage fin)
Processeur			
Nombre de processeurs virtuels (vCPU)	2	64	192
Module(s) de mémoire			
Mémoire totale du système (Gio)	4.	512	2 048
Contrôleur de stockage			
Fournisseur et modèle			
Stockage local (type A)			
Nombre de disques	1	1	1
Taille de disque (Go)	5 GB	Par défaut	Par défaut
Informations sur les disques (vitesse, interface, type)	EBS	EBS	EBS
Processeur graphique			
Nombre de processeurs graphiques	Sans objet	Sans objet	8
Fournisseur et modèle	Sans objet	Sans objet	NVIDIA H100
Fonctionnalités supplémentaires	Sans objet	Sans objet	3 200 Gbit/s de bande passante réseau[1]

Instances de la solution Azure Machine Learning

Tableau 4 : Informations de configuration détaillées pour les instances Azure.

Informations de configuration	D2 v2 (ordinateurs portables)	M64 (traitement)	ND96amsr A100 v4 (inférence et réglage fin)
Nombre d'instances	20	1	8
Fournisseur de services Cloud (CSP)	Azure	Azure	Azure
Région	Est des États-Unis 2	Est des États-Unis 2	Est des États-Unis 2
Processeur			
Nombre de processeurs virtuels (vCPU)	2	64	96
Module(s) de mémoire			
Mémoire totale du système (Gio)	7	1 000	1 900
Stockage local (type A)			
Nombre de disques	1	1	1
Taille de disque (Go)	100	7 168	6 400
Informations sur les disques (vitesse, interface, type)	Temporaire	Temporaire	Temporaire
Nombre de processeurs graphiques	Sans objet	Sans objet	8
Fournisseur et modèle	Sans objet	Sans objet	NVIDIA A100
Informations complémentaires	Sans objet	Sans objet	Chaque processeur graphique dispose d'une connexion NVIDIA Mellanox HDR InfiniBand 200 Go/s dédiée[2]

Introduction

Pour fournir un exemple des coûts d'une solution d'IA, nous avons créé un scénario d'IA à l'aide du modèle open source Llama 2 13B et comparé le coût d'exécution de la charge applicative dans quatre environnements différents. Nous avons dimensionné et estimé les coûts de quatre solutions :

- Solution Dell traditionnelle sur site
- Solution sur site gérée avec une solution Dell APEX Pay-per-Use
- Solution AWS SageMaker
- Solution Microsoft Azure Machine Learning

Les solutions Dell traditionnelles et gérées sur site utilisent le même matériel. Avec le plan traditionnel, l'entreprise achète le matériel à l'avance ; avec la solution Dell APEX Pay-Per-Use, Dell installe le matériel dans le datacenter du client et facture l'entreprise mensuellement en fonction de la capacité « engagée » et de la capacité tampon.

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 2 13B, car il s'agit d'un modèle open source largement disponible, et nous avons conçu notre scénario autour d'une seule charge applicative d'IA relativement petite. 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 avons dimensionné le matériel pour chaque solution en fonction d'hypothèses concernant les heures de travail et les capacités matérielles nécessaires. Nous avons utilisé des sources publiques pour ces recherches. Nous avons utilisé des calculateurs de prix en ligne pour AWS SageMaker et Azure Machine Learning et avons demandé et reçu des devis de Dell Technologies pour les coûts en utilisant le Dell Recommended Pricing pour les deux solutions Dell. Nous n'avons effectué aucun test pratique sur les solutions de ce document.

Nos conclusions

Dans le rapport principal, nous présentons les comparaisons normalisées de chacune des deux solutions Dell sur site par rapport aux solutions Cloud AWS SageMaker et Azure Machine Learning. Les Tableaux 5 et 6 indiquent la base des coûts de ces normalisations. La valeur normalisée est le résultat de la division de chaque valeur par le coût de la solution sur site Dell indiquée dans le tableau. Le Tableau 5 montre que les solutions Cloud coûtent jusqu'à 2,88 fois plus cher que la solution sur site traditionnelle Dell sur trois ans. La ligne du seuil de rentabilité montre que 3 ans d'utilisation de la solution Dell traditionnelle sur site coûtent à peu près le même montant qu'une seule année d'utilisation des deux solutions Cloud.

Tableau 5 : Coût total de propriété normalisé sur trois ans pour la solution traditionnelle sur site Dell par rapport aux solutions SageMaker et Azure Machine Learning.

	Coûts de la solution traditionnelle sur site Dell de 3 ans.	Engagement AWS SageMaker de 3 ans	Engagement Azure Machine Learning de 3 ans
Total	817 880,00 $	2 357 549,00 $	2 231 805,00 $
Normalisé	1	2,88	2,72
Seuil de rentabilité en mois pour la solution Dell par rapport aux solutions Cloud	Sans objet	12,5	13,3

Le Tableau 6 montre que les solutions Cloud coûtent jusqu'à 3,81 fois le prix d'une solution Dell APEX Pay-Per-Use sur 3 ans et que 3 ans de solution sur site traditionnelle Dell coûtent moins cher que le prix d'une seule année des deux solutions Cloud.

Tableau 6 : Coût total de propriété normalisé sur 3 ans pour une solution Dell APEX Pay-Per-Use par rapport aux solutions SageMaker et Azure Machine Learning.

	Coût de la solution Dell APEX Pay-Per-Use sur 3 ans	Engagement AWS SageMaker de 3 ans	Engagement Azure Machine Learning de 3 ans
Total	618 648,00 $	2 357 549,00 $	2 231 805,00 $
Normalisé	1	3,81	3,60
Seuil de rentabilité en mois pour la solution Dell par rapport aux solutions Cloud	Sans objet	9,5	10

Considérations relatives au stockage

Nous n'avons pas inclus les coûts de stockage au-delà de ce qui est nécessaire pour que les serveurs ou les instances effectuent leurs tâches.

Les solutions Dell sur site comprennent 106,88 To de stockage :

- 5,76 To de capacité SSD sur chacun des trois nœuds principaux PowerEdge R660
- 44,8 To de capacité SSD sur chacun des deux nœuds de travail de processeur graphique PowerEdge XE96-80.

La gestion des clusters et les tâches d'ordinateurs portables partagent le stockage sur les nœuds principaux. Les tâches de traitement, de réglage fin du modèle et d'inférence partagent l'espace de stockage sur les nœuds de travail de processeur graphique. Nous avons provisionné les clusters Dell PowerEdge avec un stockage supplémentaire par rapport aux solutions Cloud, afin de garantir de la place pour les tâches de gestion sur les nœuds principaux et pour une éventuelle croissance.

La solution AWS SageMaker comprend un total de 63,444 To d'espace de stockage :

- 7 000 Go de stockage EBS gp2 achetés pour chacune des deux instances de traitement ml.r5.16xlarge
- 8 disques SSD NVMe de 3084 Go pour chacune des instances ml.p5.48xlarge
- 5 Go de stockage temporaire EBS pour chaque instance d'ordinateur portable incluse avec l'instance

La solution Azure Machine Learning comprenait un stockage temporaire total de 64,82 To :

- 7 168 Gio de stockage temporaire pour l'instance de traitement M64
- 64 000 Go d'espace de stockage temporaire pour chacun des huit serveurs ND96amsr A100 v4.
- 100 Gio de stockage temporaire pour chaque instance d'ordinateur portable incluse avec l'instance

Les besoins en stockage varient pour les instances d'ordinateurs portables. Ils ne sont généralement pas très importants pour ce type de charge applicative. Nous avons donc choisi de laisser les instances Cloud avec le stockage fourni par défaut. Nous avons inclus les coûts de transfert de données pour le transfert de données EBS dans AWS SageMaker et pour le transfert de stockage Blob dans les solutions Azure Machine Learning. Nous n'avons pas inclus les coûts de transfert de données pour les solutions Dell sur site, qui utiliseraient des disques SSD intégrés.

Heures d'utilisation

Nous avons dimensionné les solutions en fonction des estimations suivantes en heures par mois pour les ordinateurs portables, le traitement des données, le réglage fin du modèle et l'utilisation des inférences. Nous avons utilisé ces heures pour calculer les heures d'utilisation des instances pour les solutions Cloud et pour dimensionner ces instances et les serveurs pour les solutions sur site.

Nous avons dimensionné les solutions en supposant qu'il y a 22 jours de travail par mois, avec des charges applicatives configurées pour s'exécuter pendant la nuit afin d'optimiser l'utilisation. Par conséquent, chaque instance de serveur et de Cloud disposerait de 528 heures d'exécution chaque mois. (Voir le Tableau 7.)

Tableau 7 : Heures d'utilisation pour les quatre tâches.

Tâche	Heures totales par mois	Calculs d'utilisation
Ordinateur portable	3 520	Nous avons dimensionné chaque solution de sorte à prendre en charge 20 professionnels des données, avec une instance d'ordinateur portable pour chacun, 8 heures par jour, 22 jours par mois, soit un total de 176 heures par mois, c'est-à-dire 3 520 heures pour les 20 professionnels des données. La configuration minimale requise est de petites instances d'ordinateurs portables Cloud avec 2 vCPU et au moins 4 Gio de mémoire.
Traitement	1 056	Les tâches de traitement des données s'exécuteraient pendant les 528 heures de disponibilité sur deux serveurs Dell PowerEdge XE9680 pour un total de 1 056 heures d'exécution et nécessiteraient 1 056 heures d'exécution sur les instances Cloud. La configuration minimale requise pour les instances Cloud était de 64 vCPU, 1 000 Gio de mémoire et 7 000 Go de stockage. Nous avons dimensionné les serveurs Dell PowerEdge de sorte à prendre en charge ces exigences, ainsi que celles des tâches de réglage fin et d'inférence.
Réglage fin du modèle	792	L'autonomie combinée de 1 056 heures pour les tâches de réglage fin et d'inférence nécessite deux serveurs Dell PowerEdge XE9680. Toutes les tâches utilisent huit processeurs graphiques H100 et jusqu'à un demi-To de mémoire. Les solutions Cloud nécessitent le même nombre d'heures sur les instances dotées de huit processeurs graphiques H100 ou équivalents. Pour AWS, nous avons utilisé deux instances H100 ; pour Azure Machine Learning, nous avons remplacé huit instances A100.
Inférence	264	

Détails des ordinateurs portables

Les scientifiques de données auraient besoin de petits ordinateurs portables Cloud avec des instances de 2 processeurs virtuels et au moins 4 GiB de mémoire. Bien que certaines tâches d'ordinateur portable puissent fonctionner mieux avec plus de mémoire, nous avons opté pour l'instance AWS ml.t3.Medium en nous basant sur les suggestions du guide « The total cost of ownership (TCO) of Amazon SageMaker »[3], puis avons choisi une instance de taille similaire pour Azure. Pour les solutions Dell sur site, nous avons supposé un cœur de processeur équivalent et 4,3 Go de mémoire par ordinateur portable, les ordinateurs portables étant exécutés sur les 3 serveurs de gestion PowerEdge R660, avec les tâches de gestion.

Instances d'ordinateur portable AWS SageMaker et Azure Machine Learning

Pour les solutions AWS SageMaker et Azure Machine Learning, nous avons sélectionné des instances d'ordinateurs portables dotées de 2 processeurs virtuels et d'au moins 4 Gio de mémoire.

Tableau 8 : Informations de configuration clés pour les instances d'ordinateur portable AWS SageMaker et Azure Machine Learning.

Type d'instance	Instance	Nombre de processeurs virtuels par instance	Mémoire (Gio) par instance
Ordinateurs portables SageMaker	ml.t3.medium	2	4
Ordinateurs portables Azure ML	D2 v2	2	7

Ordinateurs portables des solutions sur site Dell

Les solutions Dell prennent en charge ces charges applicatives d'ordinateurs portables sur les trois nœuds principaux PowerEdge R660, qui disposent ensemble de 576 Go de mémoire et de 48 cœurs de processeur, ce qui est suffisant pour prendre en charge les tâches de gestion des clusters et les 20 charges applicatives d'ordinateurs portables. Nos hypothèses pour prendre cette décision de dimensionnement sont les suivantes :

- Les tâches de gestion occupent moins de la moitié de la capacité du processeur de ces nœuds principaux et moins de 500 Go de mémoire, avec la capacité restante disponible pour exécuter ces tâches.
- Au cours des 176 heures d'exécution de chaque charge applicative d'ordinateur portable chaque mois, un seul cœur de processeur serait utilisé, c'est-à-dire l'équivalent des 2 processeurs virtuels pour les ordinateurs portables SageMaker et Azure ML, en supposant un ratio de 1 thread : 1 processeur virtuel et 4,3 Go de mémoire correspondant aux 4 Gio définis lors du dimensionnement des ordinateurs portables en Cloud. L'ensemble des 20 ordinateurs portables fonctionnant en même temps utiliserait moins de la moitié des 48 cœurs et 15 % de la mémoire de ces systèmes.
- Toutes les tâches sur tous les systèmes ont lieu pendant moins de 72,3 % du nombre total d'heures de chaque mois, sur la base de 22 jours de travail par semaine, avec 24 heures disponibles chaque jour.

Détails du traitement

Instances de traitement AWS SageMaker et Azure Machine Learning

Les tâches de traitement s'exécutent mieux sur le processeur plutôt que sur le processeur graphique et sont parfaitement réalisées sur un ratio mémoire/cœur élevé[4]. Nous nous sommes donc concentrés sur les instances de traitement à mémoire optimisée d'AWS SageMaker et Azure Machine Learning. Notre objectif était d'atteindre au moins 64vCPU, 1 000 GiB de mémoire et 7 000 GB d''espace de stockage, ce que l'instance de traitement Azure Machine Learning M64 a presque atteint avec 64vCPU, 1 000 GiB de mémoire et 7 168 GiB d'espace de stockage temporaire. Les instances de traitement à mémoire optimisée AWS SageMaker n'offraient pas d'instance correspondant à nos spécifications. Nous avons donc choisi une paire d'instances de traitement à mémoire optimisée ml.r5.16xlarge SageMaker avec une mémoire combinée de 1 024 Gio. Ensemble, ces instances SageMaker dépassent nos besoins, avec une capacité vCPU doublée et près de deux fois supérieure à la capacité de stockage (avec le stockage EBS) de nos cibles et de l'instance Azure Machine Learning.

Tableau 9 : Informations de configuration clés pour les instances de traitement AWS SageMaker et Azure Machine Learning.

Informations de configuration de l'instance	SageMaker ml.r5.16xlarge (traitement)	Azure Machine Learning M64 (traitement)
Nombre d'instances	2	1
Fournisseur de services Cloud (CSP)	AWS	Azure
Nombre de processeurs virtuels (vCPU)	64 (128 pour 2 instances)	64
Mémoire totale du système (Gio)	512 (1 024 pour deux instances)	1 000
Nombre de disques	1 (2 pour deux instances)	1
Taille de disque (Gio)	7 000 (14 000 pour deux instances)	7 168
Informations sur les disques (vitesse, interface, type)	EBS	Temporaire

Solutions Dell sur site

Les deux nœuds de travail Dell PowerEdge XE9680 disposent d'une capacité supérieure aux charges applicatives de réglage fin et d'inférence, suffisante pour prendre en charge les charges applicatives de traitement exécutées en parallèle. Les charges applicatives de traitement s'appuient sur les processeurs, et le réglage fin du modèle et l'inférence sur les processeurs graphiques. Pour correspondre à nos spécifications cibles, les charges applicatives de traitement utiliseraient la moitié des 2 To de mémoire combinés des deux serveurs, 32 cœurs de processeur et environ 7 To des 44,8 To de stockage des deux serveurs.

Détails de réglage fin et d'inférence du modèle

Les solutions nécessitent des instances de processeur graphique ou des serveurs pour le réglage fin et l'inférence des charges applicatives avec huit processeurs graphiques NVIDIA HGX H100 ou une capacité de processeur graphique équivalente et 512 Go de mémoire.

AWS SageMaker et Azure Machine Learning

L'instance d'inférence ml.p5.48xlarge est la seule instance SageMaker qui dispose des processeurs graphiques NVIDIA HGX H100[5]. Au moment de notre étude, la meilleure machine virtuelle Azure disposait de huit machines virtuelles de processeur graphique NVIDIA A100. Nous avons donc configuré l'environnement Azure pour qu'il exécute quatre fois plus d'heures d'instance afin d'atteindre à peu près les mêmes performances que les modèles H100 dans l'environnement AWS SageMaker[6]. Les deux types d'instances ont plus de 512 Go de mémoire minimale requise.

Tableau 10 : : informations clés sur la configuration pour les instances de réglage fin et d'inférence AWS SageMaker et Azure Machine Learning.

Instances d'inférence et d'entraînement	SageMaker ml.p5.48xlarge	Machine Learning Azure ND96amsr A100 v4
Nombre d'instances	2 (1 pour le réglage fin et 1 pour l'inférence)	8 (4 pour le réglage fin et 4 pour l'inférence)
Fournisseur de services Cloud (CSP)	AWS	Azure
Nombre de processeurs virtuels (vCPU)	192	96
Mémoire totale du système (Gio)	2 048	1 900
Nombre de disques	8	1
Taille de disque (Gio)	3 084	6 400
Informations sur les disques (vitesse, interface, type)	SSD NVMe	Temporaire
Nombre de processeurs graphiques	8	8
Fournisseur et modèle	NVIDIA H100	NVIDIA A100

Solutions Dell sur site

Pour les solutions Dell sur site, nous avons dimensionné les deux nœuds worker du processeur graphique PowerEdge XE9680, de sorte à gérer les charges applicatives de réglage fin et d'inférence à l'aide des ressources du processeur graphique. Les charges applicatives de traitement sont prises en charge par les ressources de processeur et de mémoire de secours. Les deux nœuds de travail de processeur graphique PowerEdge XE9680 disposent chacun de deux processeurs Intel Xeon Platinum 8468 à 48 cœurs, d'une mémoire de 1 024 Go, de 44,8 To de stockage et d'un assemblage de 8 processeurs graphiques NVIDIA HGX H100.

Analyse des coûts

Pour les solutions Cloud, nous avons inclus le coût des licences pour les instances dont nous avions besoin pour les ordinateurs portables, le traitement, le réglage fin et l'inférence. Pour les solutions Dell sur site, nous proposons deux options de paiement pour le matériel : un modèle d'achat initial et un plan de paiement mensuel de la solution Dell APEX Pay-Per-Use. Pour les deux solutions sur site, nous avons ajouté les coûts d'administration des serveurs pour le matériel et le système d'exploitation, ainsi que les coûts du datacenter pour l'espace rack et les coûts énergétiques pour l'alimentation et le refroidissement, coûts qui ne sont pas pertinents pour les deux solutions Cloud.

Nous avons omis certains coûts, par exemple :

- Nous avons omis les coûts de travail qui pouvaient être similaires sur les quatre solutions, comme l'installation et la maintenance de logiciels Open source, le transfert et la sauvegarde de données, ainsi que les salaires des experts en science des données.
- Nous n'avons inclus les coûts logiciels pour aucune des solutions. Les instances SageMaker et Azure Machine Learning incluent certains logiciels et services, tels que les ordinateurs portables Jupyter sur les instances d'ordinateurs portables et les API de traitement avec les instances de traitement. Nous avons supposé que tous les logiciels et outils supplémentaires que les scientifiques des données installeraient sur place seraient open source. Avec les solutions Dell sur site, les scientifiques des données utiliseraient exclusivement des logiciels et des outils open source tels que Jupyter Notebook, Python et PyTorch, et les serveurs exécuteraient Ubuntu ou un autre système d'exploitation open source.
- Nous n'avons pas inclus les taxes de vente, car elles varient d'un État à l'autre et d'une entreprise à l'autre. Nous n'incluons pas les coûts de migration ni les coûts de fin de vie.

Nous nous sommes concentrés sur un cycle de vie de 3 ans pour chacune de ces solutions d'IA générative. Nous avons choisi 3 ans, car les calculateurs de tarification AWS et Azure ont plafonné les engagements à trois ans. Une période de trois ans représente également un cycle de vie raisonnable pour une solution d'IA générative sur site qui nécessite du matériel de pointe[7]. Les organisations pourraient réaffecter le matériel Dell qu'elles ont acheté par la suite, et disposeraient de 2 années restantes sur les 5 ans de services Dell ProSupport inclus pour maintenir leur productivité.

Coûts sur 3 ans pour les solutions sur site Dell

Pour les solutions sur site Dell, nous avons inclus les coûts suivants sur une période de 3 ans :

- Prix recommandé par Dell pour les serveurs et commutateurs Dell, y compris ProSupport, le service sur site le jour suivant et ProDeploy Plus pour les serveurs
- L'administrateur système doit assurer la maintenance et la sécurité du matériel et du système d'exploitation
- Coûts énergétiques liés à l'alimentation et au refroidissement
- Coûts du datacenter pour l'espace rack

Les deux solutions incluaient le même matériel et impliquaient les mêmes coûts d'administration du système, d'énergie pour l'alimentation et le refroidissement, et d'espace rack du datacenter.

Tableau 11 : Coûts sur 3 ans pour une solution sur site traditionnelle.

Solution Dell traditionnelle sur site	**Coûts sur une période de 3 ans (arrondis à l'unité supérieure)**
Matériel Dell avec 5 ans de ProSupport et ProDeploy Plus (pour les serveurs)	680 251 $
Administration système	6 531 $
Coûts énergétiques liés à l'alimentation et au refroidissement	88 978 $
Coûts du datacenter pour l'espace rack	42 120 $
Total	817 880 $

Pour la solution sur site traditionnelle, nous avons demandé un devis au service commercial Dell Technologies pour le prix recommandé de la solution Dell sur site. Le 20 mars 2024, Dell a indiqué le prix d'achat du matériel à 680 251 $.

La proposition commerciale incluait 5 ans de ProSupport et de service sur site le jour ouvré suivant pour les serveurs et les commutateurs, et ProDeploy Plus pour les serveurs. Dell définit le prix recommandé comme un point de départ pour les acheteurs potentiels. Il représente le coût immédiatement accessible aux entreprises, même si elles ne sont pas des clients existants, et fonctionne essentiellement comme un prix de vente suggéré pour leurs produits.

Tableau 12 : Coûts sur 3 ans pour une solution Dell APEX Pay-Per-Use.

Solution Dell APEX Pay-Per-Use	**Coûts sur une période de 3 ans (arrondis à l'unité supérieure)**
Matériel Dell avec 5 ans de ProSupport et ProDeploy Plus (pour les serveurs)	481 019 $
Administration système	6 531 $
Coûts énergétiques liés à l'alimentation et au refroidissement	88 978 $
Coûts d'espace rack du datacenter	42 120 $
Total	618 648 $

Solution Dell APEX Pay-Per-Use sur site

En se basant sur ce prix recommandé, Dell Technologies a fourni une estimation du coût sur 3 ans d'une solution Dell APEX Pay-Per-Use de 481 019 $. Dell a envoyé à PT un devis pour la solution Dell APEX Pay-per-Use pour le même matériel que ci-dessus, pour une durée de 36 mois, avec un engagement à 75 % de capacité. Nous avons reçu cette proposition commerciale le 2 avril 2024. L'estimation de la capacité de 75 % était l'option de capacité la plus proche qui couvrirait les 528 heures de disponibilité que nous avons dimensionnées pour que les solutions fonctionnent.

Administration système

Les administrateurs de serveur surveillent et garantissent les performances, la disponibilité, les fonctionnalités et la sécurité du matériel et du système d'exploitation et, dans ce cas, installent le système d'exploitation. Ces services sont requis par la solution sur site, mais pas par les solutions Cloud, car ils sont inclus dans leur contrat de service. Nous avons maintenu ces estimations de temps et de coûts à un niveau bas pour la solution sur site, car nous supposons qu'elles sont principalement automatisées et que ProSupport for Infrastructure décharge certaines des tâches de surveillance et de réparation physique des administrateurs système sur site.

Nous avons estimé un coût de 6 531,00 USD sur trois ans pour cette administration de serveur, basé sur la rémunération totale d'un administrateur système de niveau intermédiaire[8] qui est en mesure de gérer 300 serveurs avec les commutateurs et systèmes d'exploitation associés à l'aide d'outils et de processus automatisés et qui bénéficie de l'aide des services ProSupport et ProDeploy Plus.

Dell ProDeploy Plus for Infrastructure

Nous n'avons pas inclus d'estimation distincte pour le déploiement. Nous nous sommes plutôt appuyés sur Dell ProDeploy Plus for Infrastructure, un service que nous avons inclus dans la proposition commerciale matérielle, pour fournir le déploiement matériel logiciel sur site[9]. Un rapport Principled Technologies indique que ProDeploy Plus for Infrastructure peut « faire gagner un temps précieux aux administrateurs internes en faisant appel à un ingénieur certifié Dell technologies pour l'installation et la configuration d'un système Dell »[10].

Ce service peut ne pas couvrir certaines tâches de planification, de déballage et de mise en rack des commutateurs, ou encore d'installation du système d'exploitation. Ces tâches prendraient peu de temps et sont incluses dans l'estimation du temps d'administration du système pour maintenir et sécuriser la solution.

Dell ProSupport for Infrastructure

Nous avons inclus 5 ans de Dell ProSupport et Next Day Onsite Service. La durée de support de 5 ans est plus longue que la période de 3 ans de notre analyse, mais offre à l'entreprise acheteuse la valeur ajoutée d'un cycle de vie plus long pour le matériel Dell.

Coûts énergétiques liés à l'alimentation et au refroidissement

Les solutions Cloud incluaient les coûts énergétiques de l'alimentation et du refroidissement dans leurs prix. Pour les solutions sur site, nous avons estimé les coûts d'alimentation et de refroidissement sur 3 ans à l'aide de l'outil Dell Enterprise Infrastructure Planning[11]. Pour obtenir une estimation, nous avons saisi les spécifications des serveurs et commutateurs inclus dans la proposition commerciale Dell Technologies. Nous avons fourni deux autres données qui ont affecté les calculs :

- 1,58 multiplicateur de l'efficacité de l'utilisation de l'énergie (PUE) des coûts d'alimentation pour obtenir les coûts combinés d'alimentation et de refroidissement. Le PUE était la moyenne du secteur en 2023, selon l'Uptime Institute, une organisation qui enquête et suit les coûts des datacenters[12].
- 12,74 cents par kilowatt-heure de coût énergétique basé sur les données de l'Agence d'information sur l'énergie (EIA) américaine sur le prix moyen de détail de l'électricité pour le secteur commercial en 2023[13].

Nous avons calculé les coûts d'alimentation et de refroidissement séparément pour les appareils exécutant des charges applicatives inactives et de calcul. Nous avons évalué les résultats en fonction des 528 heures d'exécution (environ 72,3 % d'un mois moyen) que nous avons dimensionnées pour que les solutions fonctionnent.

Tableau 13 : Consommation énergétique sur 3 ans pour l'alimentation et le refroidissement[14].

Charges applicatives	Coût énergétique sur 3 ans pour une solution sur site	Facteur de pondération	Coût énergétique pondéré pour l'alimentation et le refroidissement
Calcul	114 439,69 $	72,3 %	82 739,90 $
Inactif	22 516,69 $	27,7 %	6 237,12 $
Coût énergétique pondéré sur 3 ans pour l'alimentation et le refroidissement			88 977,02 $

Coût des racks de datacenter

L'entreprise devrait supporter des coûts supplémentaires pour héberger les serveurs et les racks dans le datacenter. Ces coûts incluent les coûts du rack, de la gestion de réseau et d'autres coûts liés aux installations du datacenter. Nous avons estimé ces coûts à 1 800 $ par rack et par mois pour les racks d'une capacité utile de 28 u[15]. Les serveurs et commutateurs de cette solution occupent 18 u, soit 65 % d'un seul de ces racks. Sur trois ans, le coût de cette utilisation est de 42 120 $.

Même si nous avons inclus le transfert de données pour les deux solutions Cloud afin d'accéder au stockage AWS S3 et au stockage Azure Block, nous n'avons pas ajouté ces coûts pour les solutions sur site, car elles utiliseraient leurs disques intégrés ou leurs baies de stockage locales pour le stockage.

Solution AWS SageMaker

Nous avons configuré la solution AWS SageMaker pour qu'elle corresponde aussi exactement que possible à la solution Dell sur site citée pour les quatre tâches que nous avons décrites précédemment : ordinateurs portables, traitement, réglage fin du modèle et inférence. AWS Pricing Calculator for SageMaker répertorie chaque tâche en tant que module de tarification distinct que vous pouvez ajouter à l'estimation. Nous avons ajouté SageMaker Studio Notebooks, SageMaker Processing, SageMaker Training et SageMaker Real-Time Inference. Pour chaque module, nous avons rempli les champs nécessaires pour déterminer le coût horaire de chaque instance choisie pour chaque tâche. Nous avons ensuite utilisé ce coût horaire pour déterminer le montant qu'un utilisateur consacrerait à exécuter chaque instance pour le nombre d'heures précalculé que nous avons déterminé en fonction des systèmes Dell. Nous avons calculé deux fois plus d'instances de traitement et, par conséquent, d'heures de traitement pour nous assurer que la capacité de traitement des autres solutions correspond à celle des autres solutions. Nous avons également ajouté le stockage EBS aux instances de traitement, car elles ne tournent pas avec le stockage en dehors du volume du système d'exploitation. Après l'application du plan d'économies sur 3 ans, nos coûts totaux se sont élevés à 2 357 549 USD pour 3 ans. Voir le Tableau 14 pour obtenir les détails complets de l'instance.

Tableau 14 : Coûts des instances de la solution SageMaker sur 3 ans.

Service	Type d'instance	Valeur de l'instance en $/heure	Temps d'exécution (heures/mois)	Coûts pour 3 ans
Ordinateurs portables SageMaker Studio	ml.t3.medium	0,02244 $	3520	2 843,60 $
Traitement SageMaker	ml.r5.16xlarge*	2,2908 $	2112	174 174,11 $
Stockage EBS traitement SageMaker (7 To par mois)				25 804,80 $
Entraînement SageMaker	ml.p5.48xlarge**	56,5340 $	792	1 611 897,41 $
Inférence SageMaker en temps réel	ml.p5.48xlarge**	56,5340 $	264	537 299,14 $
Transfert de données S3 (1 entrée et 15 sorties)				5 529,60 $
Total (arrondi à l'unité supérieure)				2 357 549 $

*Nous incluons deux instances de processeur pour obtenir 1 To de mémoire pour les tâches de traitement.
**Prix estimés en fonction du pourcentage d'engagement P5 public (non ml.p5) car le calculateur n'a pas inclus cette instance.

Solution Azure Machine Learning

Nous avons utilisé Azure Pricing Calculator pour connecter chaque type d'instance afin de déterminer le coût horaire de chaque instance dans le service Machine Learning. Étant donné que la meilleure instance de processeur graphique Azure proposée au moment de cette étude incluait huit processeurs graphiques A100, nous avons calculé les coûts d'exécution de quatre fois le nombre de ces instances de processeur graphique par tâche, afin de fournir une approximation plus précise des performances des huit processeurs graphiques H100 que les instances AWS et les serveurs Dell PowerEdge XE9 fournissent. Nous avons ensuite utilisé ce coût horaire pour déterminer le montant qu'un utilisateur consacrerait à exécuter chaque instance pour le nombre d'heures précalculé que nous avons déterminé en fonction des systèmes Dell. Après avoir appliqué la tarification d'engagement réservé Azure sur 3 ans, nos coûts totaux pour 3 ans se sont élevés à 2 231 805 $. Voir le Tableau 15 pour plus d'informations sur l'instance.

Tableau 15 : Coûts des instances de la solution Azure Machine Learning sur 3 ans.

Service	Type d'instance	Valeur de l'instance en $/heure	Temps d'exécution (heures/mois)	Coûts pour 3 ans
Ordinateurs portables Azure ML	D2 v2	0,0476 $	3 520	6 027,72 $
Traitement Azure ML	M64	1,8258 $	1 056	69 407,81 $
Entraînement Azure ML	ND96amsr A100 v4	14,1475 $	3 168	1 613 491,01 $
Inférence Azure ML en temps réel	ND96amsr A100 v4	14,1475 $	1 056	537 830,34 $
Opérations de transfert de données Azure Block Blob Storage (10 000 000)				5 047,20 $
Total (arrondi à l'unité supérieure)				2 231 805 $

1. AWS, « Get Started with P5 instances », consulté le 29 avril 2024, https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/p5-instances-started.html.
2. Microsoft Build, « NDm A100 v4-series », consulté le 29 avril 2024, https://learn.microsoft.com/en-us/azure/virtual-machines/ndm-a100-v4-series
3. Amazon, « Total Cost of Ownership of Amazon SageMaker », consulté le 29 avril 2024, https://pages.awscloud.com/rs/112-TZM-766/images/Amazon_SageMaker_TCO_uf.pdf.
4. StackOverflow, « Why should preprocessing be done on CPU rather than GPU? », consulté le 20 avril 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu et Hugging Face, « Model Memory Requirements », consulté le 29 avril 2024, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
5. AWS, « New – Amazon EC2 P5 Instances Powered by NVIDIA H100 Tensor Core GPUs for Accelerating Generative AI and HPC Applications », consulté le 29 avril 2024, https://aws.amazon.com/blogs/aws/new-amazon-ec2-p5-instances-powered-by-nvidia-h100-tensor-core-gpus-for-accelerating-generative-ai-and-hpc-applications/.
6. Comet, « Comparison of NVIDIA A100, H100 + H200 GPUs », consulté le 29 avril 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.
7. The Jérusalem Post, « Maximizing Efficiency: Your 2023 Guide to GPU Servers », consulté le 8 avril 2024, https://www.jpost.com/insights/article-770858.
8. Rémunération totale (salaire et avantages sociaux) de l'administrateur système II de 130 616 $ par an. Source : Salary.com, « Systems Administrator II », consulté le 25 mars 2024, https://www.salary.com/tools/salary-calculator/systems-administrator-ii-benefits.
9. Dell, « The market's most complete deployment offer », consulté le 28 mars 2024, https://www.delltechnologies.com/asset/en-us/services/deployment/briefs-summaries/prodeploy_plus_deployment_unification_ds.pdf.
10. Principled Technologies, « L'utilisation de Dell ProDeploy Plus for Infrastructure peut améliorer le temps de déploiement pour la technologie Dell », consulté le 28 mars 2024, https://www.delltechnologies.com/asset/fr-fr/products/cross-company/industry-market/principled-technologies-prodeploy-plus-for-infrastructure-services-whitepaper.pdf.
11. Dell, « Dell Enterprise Infrastructure Planning Tool », consulté le 24 mars 2024, https://dell-ui-eipt.azurewebsites.net/#/
12. Uptime Institute, « Large data centers are mostly more efficient, analysis confirms », consulté le 29 mars 2024, https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/.
13. EIA, « Electricity Data Browser », consulté le 29 mars 2024, https://www.eia.gov/electricity/data/browser/#/topic/7?agg=0,1&geo=g&endsec=vg&linechart=ELEC.PRICE.US-ALL.A~ELEC.PRICE.US-RES.A~ELEC.PRICE.US-COM.A~ELEC.PRICE.US-IND.A&columnchart=ELEC.PRICE.US-ALL.A~ELEC.PRICE.US-RES.A~ELEC.PRICE.US-COM.A~ELEC.PRICE.US-IND.A&map=ELEC.PRICE.US-ALL.A&freq=A&ctype=linechart&ltype=pin&rtype=s&maptype=0&rse=0&pin=.
14. Dell, « Dell Enterprise Infrastructure Planning Tool », consulté le 25 mars 2024, https://dell-ui-eipt.azurewebsites.net/#/.
15. VMware par le partenaire Broadcom, Softchoice, utilise ce coût de rack dans une analyse de coût total de propriété comparant les coûts de l'exécution de VMware sur site ou dans le Cloud. Source : SoftChoice, « VMware Cloud on AWS », consulté le 24 mars 2024, https://www.softchoice.com/technology-partners/vmware/cloud-on-aws-tco-calculator.

▶ Consultez la version en anglais d'origine de ce rapport à l'adresse https://facts.pt/9PHKEUe

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

Principled Technologies®

Facts matter.®