Enterprise Strategy Group by TechTarget

RÉSUMÉ ÉCONOMIQUE

AVRIL 2024

Optimiser le retour sur investissement de l'IA : l'inférence sur site avec Dell Technologies peut être 75 % plus avantageuse sur le plan économique que le Cloud public

Aviv Kaufmann, Practice Director et Principal Validation Analyst

Cliquez ici pour lire l'intégralité du livre blanc économique.

Économies attendues avec l'utilisation de l'infrastructure Dell Technologies pour l'inférence LLM

38 % à 48 % plus rentable qu'une solution IaaS pour l'inférence de LLM plus petits (7 milliards de paramètres)

69 % à 75 % plus rentable qu'une solution IaaS pour l'inférence de LLM plus volumineux (70 milliards de paramètres)

Jusqu'à 88 % plus rentable que des services d'API pour l'inférence de LLM plus volumineux (70 milliards de paramètres)

Résumé : Enterprise Strategy Group, une division de TechTarget, a modélisé le coût prévisionnel associé à l'inférence de grands modèles de langage (LLM) sur une infrastructure Dell Technologies sur site, et l'a comparé à l'utilisation d'une infrastructure de Cloud public native as-a-Service (IaaS) ou du service de LLM OpenAI GPT-4 Turbo fourni via une API. Il est apparu que Dell Technologies était capable d'assurer une inférence LLM sur site à un coût jusqu'à 75 % inférieur à celui d'un Cloud public en mode natif et jusqu'à 88 % inférieur au coût associé à l'utilisation de services d'API.

Défis pour les entreprises

L'IA générative (GenAI) et les LLM, qui exploitent les données spécifiques de l'entreprise et d'autres propriétés intellectuelles pour automatiser la génération de contenu, répondre à des questions et rendre les informations facilement accessibles aux décideurs, sont de plus en plus massivement adoptés parmi les organisations. Un LLM peut se révéler coûteux et complexe à développer de A à Z, mais les organisations ont la possibilité d'augmenter, d'affiner et de personnaliser facilement les LLM Open Source existants pour les adapter à leurs besoins. Les organisations peuvent accéder à des services basés sur des API, tels qu'OpenAI GPT, mais les coûts liés à l'inférence (c'est-à-dire l'interrogation) peuvent rapidement grimper. Enterprise Strategy Group a constaté que, pour développer et utiliser une IA générative prise en charge par un LLM, les organisations avaient tendance à privilégier l'utilisation d'un LLM Open Source associé au développement d'une solution d'IA générative en interne.[1] Les organisations peuvent créer et contrôler leur propre solution d'inférence LLM sur des serveurs d'entreprise optimisés par de puissants processeurs graphiques, ou en utilisant des instances Cloud équivalentes alimentées par un processeur graphique et complétées par une plateforme d'apprentissage automatique (comme la plateforme NVIDIA AI Enterprise, par exemple) exécutant des LLM Open Source tels que Mistral ou Llama.

La solution : choisir Dell Technologies pour l'inférence LLM

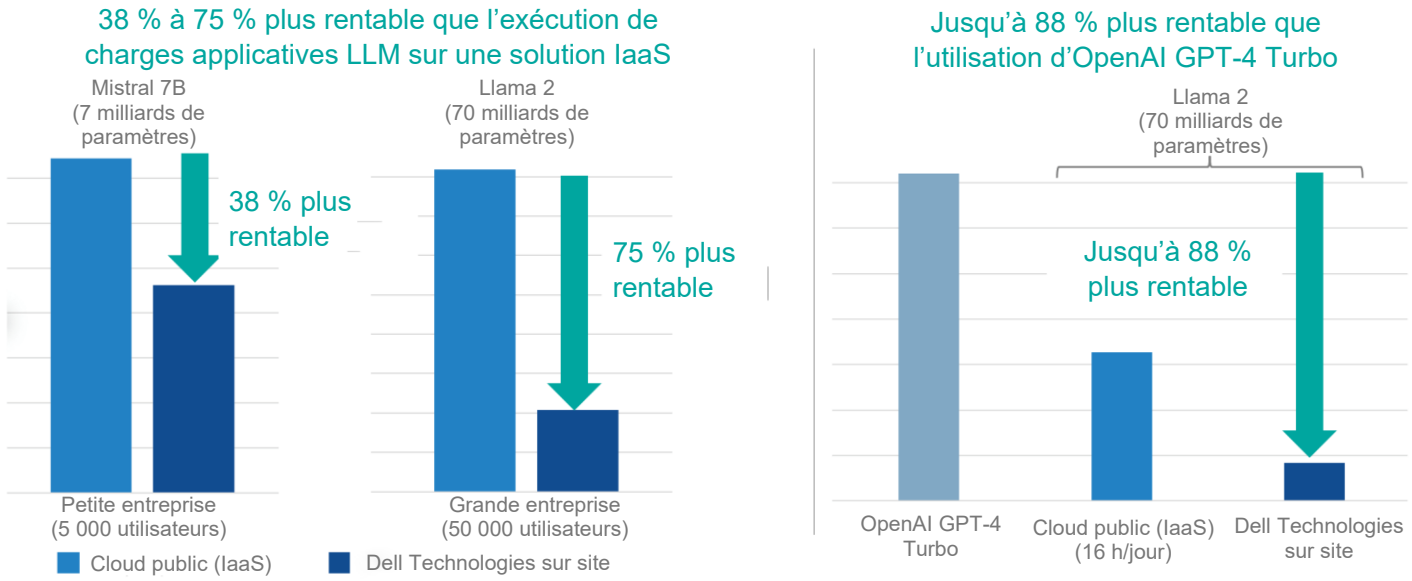
Dell Technologies offre aux organisations la possibilité d'intégrer l'IA à leurs données, où qu'elles se trouvent : en périphérie, sur site, en colocation, dans le datacenter, dans les environnements de Cloud public ou encore sur l'appareil lui-même. Dell simplifie et accélère la transition vers l'IA générative, en produisant de meilleurs résultats adaptés aux besoins de l'entreprise et en protégeant les données propriétaires de manière sécurisée et durable. En plus de l'infrastructure matérielle et logicielle, Dell propose un écosystème robuste de partenaires et de services pour aider les organisations qui débutent leur transition vers l'IA générative ou qui cherchent à aller plus loin, en fournissant des solutions complètes qui offrent une flexibilité optimale, pour aujourd'hui comme pour demain. Pour en savoir plus sur la façon dont Dell peut vous aider, rendez-vous sur sa page Web dédiée à l'IA générative. Avec Dell APEX, les organisations peuvent également opter pour des solutions d'IA générative sur abonnement et les optimiser pour les cas d'utilisation multicloud.

[1] Source : rapport d'étude Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, août 2023.

Points clés de l'analyse économique

Enterprise Strategy Group a modélisé les coûts prévisionnels associés à l'inférence de LLM textuels à 7 milliards et 70 milliards de paramètres utilisant l'approche RAG (Retrieval-Augmented Generation) pour des organisations de différentes tailles sur du matériel Dell Technology, et les a comparés à l'utilisation d'une solution IaaS de type Cloud public via des instances Amazon EC2 et à l'utilisation de l'API OpenAI GPT-4 Turbo. Les configurations des serveurs Dell Technologies et des processeurs graphiques NVIDIA H100 ont été dimensionnées compte tenu des résultats des tests d'inférence de référence afin de s'assurer qu'elles respecteraient les paramètres de simultanéité et de temps de réponse utilisables. Nous avons ensuite pris en compte les coûts associés au matériel, aux logiciels, au support et aux services, aux licences NVIDIA AI Enterprise, à l'alimentation et au refroidissement, ainsi qu'à l'administration de l'infrastructure et de la plateforme d'IA. Nous avons dimensionné et tarifé les instances de Cloud public natives avec des capacités de processeur, de mémoire et de processeur graphique H100 aussi équivalentes que possible, en tenant compte des remises sur les réservations, en supposant 16 h d'opérations par jour du lundi au vendredi avec des licences NVIDIA AI Enterprise à l'heure et en prenant en compte les avantages d'administration du Cloud. Enfin, nous avons modélisé les coûts prévisionnels engagés par les utilisateurs pour accéder à l'API OpenAI GPT-4 Turbo sur différentes fréquences d'inférences générées par les utilisateurs. Les résultats ont révélé que Dell Technologies pouvait fournir une capacité d'inférence à un coût jusqu'à quatre fois plus avantageux que celui du Cloud public pour ces cas d'utilisation sur une période de trois ans.

Figure 1. Coûts modélisés par Enterprise Strategy Group pour gérer l'inférence LLM



Source : Enterprise Strategy Group, une division de TechTarget, Inc.

Conclusion

Enterprise Strategy Group recommande vivement aux entreprises qui mettent en œuvre des LLM pour optimiser leur organisation d'envisager de tirer parti des technologies économiques et des services d'expertise fournis par Dell Technologies afin de garantir un résultat positif et d'accélérer leurs initiatives d'IA générative.

Cliquez ici pour lire l'intégralité du livre blanc économique.

À propos d'Enterprise Strategy Group
Enterprise Strategy Group de TechTarget fournit des informations ciblées et exploitables sur le marché, des recherches sur la demande, des services consultatifs d'analystes, des conseils en matière de stratégie GTM, des validations de solutions et du contenu personnalisé pour soutenir l'achat et la vente de technologies d'entreprise.

contact@esg-global.com
www.esg-global.com