Enterprise Strategy Group by TechTarget

EKONOMISK SAMMANFATTNING

APRIL 2024

Maximera avkastning på investering för AI: Lokal inferens med Dell Technologies kan vara 75 % mer kostnadseffektivt än offentliga moln

Aviv Kaufmann, verksamhetschef och huvudansvarig valideringsanalytiker

Klicka för att läsa hela det ekonomiska informationsdokumentet.

Förväntade besparingar vid LLM-inferens med Dell Technologies-infrastruktur

38 % till 48 % mer kostnadseffektivt än IaaS för inferens av mindre LLM-modeller (7 md parametrar)

69 % till 75 % mer kostnadseffektivt än IaaS för inferens av större LLM-modeller (70 md parametrar)

Upp till 88 % mer kostnadseffektivt än API-tjänster för inferens av större LLM-modeller (70 md parametrar)

Sammanfattning: TechTargets Enterprise Strategy Group modellerade och jämförde de förväntade kostnaderna för inferens med stora språkmodeller (LLM) på lokal Dell Technologies-infrastruktur jämfört med att använda infrastructure-as-a-service (IaaS) med inbyggda offentliga moln eller OpenAI GPT-4 Turbo LLM-modelltjänsten via ett API. Vi konstaterade att Dell Technologies kunde tillhandahålla lokal LLM-inferens upp till 75 % mer kostnadseffektivt än med inbyggda offentliga moln och upp till 88 % mer kostnadseffektivt än med API-tjänster.

Utmaningar för företag

Organisationer börjar satsa på generativ AI (GenAI) och LLM:er som utnyttjar företagsspecifika data och annan immateriell egendom för att automatisera innehållsgenerering, besvara frågor och göra insikter lättillgängliga för beslutsfattare. En LLM kan vara kostsam och komplex att utveckla från grunden, men organisationer kan enkelt utöka, finjustera och anpassa befintliga LLM:er med öppen källkod för att uppfylla sina behov. Organisationer har åtkomst till API-baserade tjänster som OpenAI GPT, men kostnaderna för inferens (dvs. frågeställningar) kan snabbt bli höga. Enterprise Strategy Group fann att den mest populära strategin för organisationer att utveckla och använda GenAI med stöd av en LLM var att använda en LLM med öppen källkod och utveckla en GenAI-lösning internt.[1] Organisationer kan bygga och styra sin egen LLM-inferenslösning på kraftfulla GPU-aktiverade företagsservrar eller motsvarande GPU-aktiverade molninstanser och en maskininlärningsplattform som NVIDIAs AI Enterprise som kör LLM:er med öppen källkod som Mistral eller Llama.

Lösningen – LLM-inferens med Dell Technologies

Dell Technologies gör det möjligt för organisationer att använda AI till sina data, oavsett var de finns, inklusive lokal kantanvändning, samlokalisering, datacenter, offentliga molnmiljöer och på själva enheten. Dell förenklar och påskyndar organisationers GenAI-resor, skapar bättre resultat som är anpassade till företagets behov och skyddar äganderättsskyddade data på ett säkert och hållbart sätt. Förutom infrastruktur för hårdvara och mjukvara erbjuder Dell ett robust ekosystem av partner och tjänster för att hjälpa organisationer, oavsett om de precis har börjat eller skalar upp sin GenAI-resa, med heltäckande lösningar som levererar ultimat flexibilitet nu och i framtiden. Om du vill veta mer om hur Dell kan hjälpa till kan du besöka deras GenAI-webbsida. Med Dell APEX kan organisationer dessutom prenumerera på GenAI-lösningar och optimera dem för flera moln.

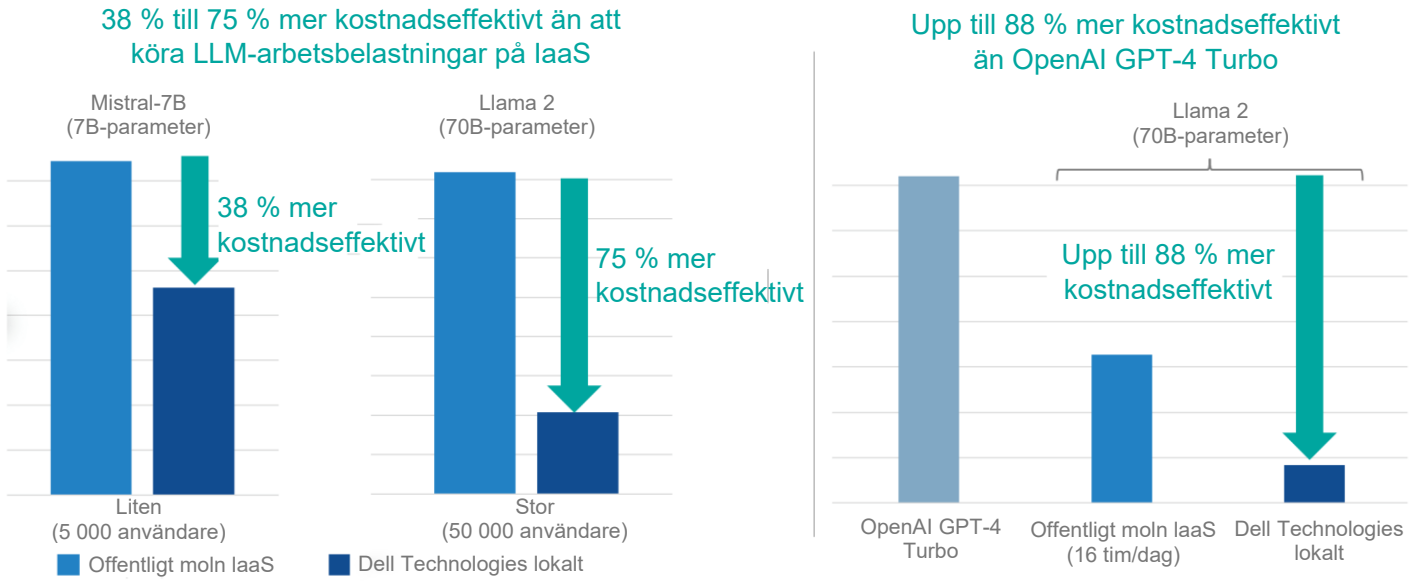
[1] Källa: Forskningsrapport från Enterprise Strategy Group, *Beyond the GenAI Hype: Real-World Investments, Use Cases, and Concerns*, augusti 2023.

Den här ekonomiska sammanfattningen från Enterprise Strategy Group har sammanställts på uppdrag av Dell Technologies och distribueras på licens från TechTarget, Inc.

Höjdpunkter i den ekonomiska analysen

Enterprise Strategy Group modellerade och jämförde de förväntade kostnaderna för att tillhandahålla inferens för textbaserade LLM-modeller med 7 md och 70 md parametrar med hjälp av RAG (retrieval-augmented generation) för organisationer av olika storlek på Dell Technologies-hårdvara, IaaS i offentliga moln på Amazon EC2-instanser och med hjälp av OpenAI GPT-4 Turbo API. Dell Technologies servrar och NVIDIA H100 GPU-konfigurationerna anpassades utifrån resultaten av baslinjetesterna för inferens för att säkerställa att de skulle uppfylla användbara parametrar för samtidighet och svarstid. Vi beaktade sedan de tillhörande kostnaderna för hårdvara, mjukvara, support och tjänster, NVIDIA AI Enterprise-licenser, strömförsörjning och nedkylning samt administration av infrastruktur och AI-plattform. Vi anpassade och prissatte de ursprungliga offentliga molninstanserna med så likvärdiga processorer, minne och H100 GPU-egenskaper som möjligt, med hänsyn till bokningsrabatter, 16-timmarsdrift från måndag till fredag, NVIDIA AI Enterprise-licensiering per timme och molnadministrativa fördelar. Slutligen modellerade vi de förväntade kostnaderna för användare att få åtkomst till OpenAI GPT-4 Turbo API över olika användargenererade inferensfrekvenser. Resultaten visade att Dell Technologies kan tillhandahålla upp till fyra gånger mer kostnadseffektiva inferenser jämfört med offentliga moln för dessa användningsfall under en treårsperiod.

Bild 1 Enterprise Strategy Groups modellerade kostnad för att hantera LLM-inferens



Källa: Enterprise Strategy Group, a division of TechTarget, Inc.

Sammanfattning

Enterprise Strategy Group rekommenderar starkt att företag som implementerar LLM för att stärka sina organisationer överväger att dra nytta av den kostnadseffektiva teknik och de kunniga tjänster som Dell Technologies tillhandahåller för att säkerställa ett framgångsrikt resultat och påskynda sina GenAI-initiativ.

Klicka för att läsa hela det ekonomiska informationsdokumentet.

Om Enterprise Strategy Group
TechTarget's Enterprise Strategy Group erbjuder fokuserad och konstruktiv marknadsinformation, undersökningar på efterfrågesidan, analytikerrådgivningstjänster, GTM-strategivägledning, lösningsvalideringar och anpassat innehåll som stödjer köp och försäljning av företagsteknik.

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