Investera i GenAI: Kostnads-nyttoanalys av Dells lokala distributioner jämfört med liknande AWS- och Azure-distributioner

I teknikvärlden är generativ AI (GenAI) en ny milstolpe. Företag över hela världen börjar nu utforska hur GenAI kan hjälpa dem nå sina affärsmål. Men de stöter också på många hinder längs vägen när de implementerar en GenAI-lösning som kan bemöta just deras behov. En av de största utmaningarna är att avgöra exakt hur mycket en GenAI-lösning i rätt storlek kostar och om den ska driftsättas lokalt eller i molnet. Som Gartner-analytikern Frances Karamouzis konstaterar: ”Kostnaden är ett av de största hindren för att lyckas med AI och generativ AI. Mer än hälften av alla företag ger upp på grund av misstag i att uppskatta och beräkna kostnader.”[1]

Vi ville ge företag en rimlig utgångspunkt så att de kunde se den totala kostnaden för att driftsätta och hantera GenAI-arbetsbelastningar, inklusive modellfinjustering och inferens. Därför tittade vi på de ungefärliga kostnaderna under tre år för två lokala lösningar med Dell™ som använder PowerEdge™ R660 och PowerEdge XE9680, en traditionell lösning och en prenumerationsbaserad Dell APEX Pay-per-Use-lösning, och jämförbara lösningar som Amazon Web Services (AWS) SageMaker och Microsoft Azure Machine Learning. Enligt våra beräkningar var Dell APEX Pay-per-Use-lösningen den mest kostnadseffektiva av de 3-åriga lösningar vi jämförde. De konkurrerande molnlösningarna från AWS och Azure kostar upp till 3,81 gånger så mycket som den prenumerationsbaserade Dell APEX Pay-per-Use-lösningen. Jämfört med en lokal Dell-lösning skulle AWS- och Azure-molnlösningarna som vi prissatte kosta upp till 2,88 gånger så mycket. Fortsätt läsa för att se hur GenAI kan hjälpa ditt företag och hur vi beräknade våra resultat för total ägandekostnad (TCO).

Betala mindre för GenAI med en Dell APEX Pay-per-Use-lösning och en lokal GenAI-lösning från Dell

Konkurrerande AWS- och Azure-molnlösningar kostar mer:

Upp till 3,81 gånger kostnaden för en Dell APEX Pay-per-Use-lösning

Upp till 2,88 gånger kostnaden för en traditionell och lokal Dell-lösning

Fördelarna med att använda förtränade modeller för generativ AI

AI-modeller (artificiell intelligens) är system som syftar till att efterlikna aspekter av mänsklig intelligens eller mänskligt beteende. Både företag och privatpersoner använder GenAI-verktyg (som ChatGPT och DALL-E) för att generera innehåll: text, ljud, videor, bilder, kod och simuleringar samt mer komplexa utdata som anpassat marknadsföringsinnehåll, anpassade program och mjukvara.[2] Ett företag som vill integrera generativ AI i sin verksamhet kan välja mellan att använda en förtränad modell eller att skapa en egen modell från grunden. En förtränad modell som Llama 2 är redan tränad med en grundläggande datauppsättning för modellens avsedda användning. Det innebär att företag kan finjustera modellen med hjälp av specifika data som drar igång processen med att skapa en modell som är anpassad till deras data och användningsfall.[3] Enligt en källa kan användning av förtränade modeller korta ner tiden det tar att ta fram en funktionell AI-modell med upp till ett år och spara hundratusentals dollar.[4]

Scenario för total ägandekostnad och lösningsöversikt

Nya arbetsbelastningar innebär ofta nya investeringar. Många AI-arbetsbelastningar kräver högpresterande komponenter utöver de stora mängder lagring som redan innehåller dina data. Om man vill ta reda på hur man implementerar AI-arbetsbelastningar bör man balansera säkerhet, tid, prestanda, skalbarhet, användarvänlighet och kostnad. Vi skapade ett AI-scenario med hjälp av Llama 2 13B-modellen med öppen källkod för att ge en uppfattning om hur mycket AI-lösningar kostar. Här jämförde vi kostnaden för att köra arbetsbelastningen i fyra olika miljöer. Vårt scenario inkluderade fyra specifika uppgifter i en GenAI-arbetsbelastning: dataexperter som kodar och utför andra uppgifter, databearbetningsuppgifter, finjustering av modeller och inferensuppgifter. Dessa uppgifter kombineras för att hålla modellen korrekt och uppdaterad med nya företagsgenererade data för att på så sätt ge optimala utdata från modellen. I tabell 1 visas specifikationerna på hög nivå för de fyra miljöer vi undersökte. Obs! Vi slutförde all forskning och prissättning den 29 mars 2024 med priser som kan ha ändrats efter detta datum.

Tabell 1: Lösningsinformation för jämförelse av total ägandekostnad.

Uppgift	Server/instans	GPU:er per server/instans	Ytterligare inköp
Traditionell lokal lösning			
Klusterhantering	3x PowerEdge R660	e.t.	2x PowerSwitch S5232-ON-nätverksinfrastruktur och 1x PowerSwitch N3200-ON OOB-hantering
Bärbara datorer			
Databehandling	2x PowerEdge XE9680	8x NVIDIA H100	
Modellfinjustering			
Inferens			
Lokal Dell APEX Pay-per-Use-lösning			
Klusterhantering	3x PowerEdge R660	e.t.	2x PowerSwitch S5232-ON-nätverksinfrastruktur och 1x PowerSwitch N3200-ON OOB-hantering
Bärbara datorer			
Databehandling	2x PowerEdge XE9680	8x NVIDIA H100	
Modellfinjustering			
Inferens			

Uppgift	Server/instans	GPU:er per server/instans	Ytterligare inköp
AWS SageMaker-lösning			
Klusterhantering	e.t.	e.t.	7 TB EBS-lagring per månad för ml.r5.16xlarge-instanser och 1 TB in och 15 TB ut för S3-dataöverföring
Bärbara datorer	20x ml.t3.medium	e.t.	
Databehandling	2x ml.r5.16xlarge	e.t.	
Modellfinjustering	ml.p5.48xlarge	8x NVIDIA H100	
Inferens	ml.p5.48xlarge	8x NVIDIA H100	
Azure Machine Learning-lösning			
Klusterhantering	e.t.	e.t.	10 000 000 dataöverföringsåtgärder för Azure Block Blob Storage
Bärbara datorer	20x D2 v2	Ej tillämpligt	
Databehandling	M64	e.t.	
Modellfinjustering	4x ND96amsr A100 v4	8x NVIDIA A100	
Inferens	4x ND96amsr A100 v4	8x NVIDIA A100	

Om du vill ha information om exakta specifikationer för de lösningar vi jämförde kan du se vetenskapen bakom rapporten.

För den här analysen försökte vi skapa ett brett tillämpligt exempelscenario för att uppskatta kostnadsskillnader i olika miljöer. Vi valde Llama 2 13B GenAI-modellen eftersom den är en allmänt tillgänglig modell med öppen källkod. Vi inkluderade kostnader för dataexperternas bärbara datorer som används för utveckling av maskininlärning, databearbetningsuppgifter, kontinuerlig modellfinjustering och realtidsinferens. Vi tog inte med kostnader för lagring utöver vad servrar eller instanser behövde för att utföra sina uppgifter.

För lokala Dell-lösningar antog vi att utvecklarnas bärbara datorer och klusterhanteringsuppgifter skulle utföras på Dell PowerEdge R660-klustret, medan bearbetnings-, finjusterings- och inferensuppgifter skulle utföras på Dell PowerEdge XE9680-klustret.

För molnlösningarna valde vi instanser som passade en viss aktivitets behov: instanser för bärbara datorer var mycket små men däremot gav vi bearbetningsinstanser en avsevärd mängd minne. Eftersom offentliga molntjänster startar upp en ny instans för varje uppgift, skulle var och en av dessa uppgifter ha en dedikerad GPU med åtta instanser för körningen. Därför beräknade vi antalet uppgifter som PowerEdge XE9680-servrarna kunde utföra samtidigt som samma förhållande per uppgift för GPU:n bibehålls. Vi lade också till en uppskattning för kostnaderna för dataöverföring till och från molnleverantörens objektlagring för att ta med kostnaden för att flytta data genom molnet.

I syfte att ta hänsyn till olika verksamhetssituationer och göra en rättvis jämförelse gjorde vi följande antaganden:

- Alla kostnader är exklusive moms eftersom specifika priser varierar beroende på plats.
- All programvara är öppen källkod, med licenser som tillåter kommersiell användning.
- Vi exkluderar hanteringskostnader för molnlösningarna. För lokala lösningar tog vi hänsyn till de löpande kostnader som krävs för systemadministration för att underhålla hårdvara och stödja dataexperterna.
- Vad gäller de lokala lösningarna tog vi hänsyn till kostnader för fysiskt datacenterutrymme, strömförsörjning och nedkylning. Vi la sedan till dessa kostnader till instanskostnader för molnlösningarna.

Mer information om våra antaganden och beräkningar finns i vetenskapen bakom rapporten.

Jämförelse av kostnaderna för GenAI: lokala Dell-lösningar med molnlösningar

Antaganden för GenAI-kostnadsjämförelser

- Vi antog att det är 22 arbetsdagar i varje månad, med arbetsbelastningar inställda på att köras över natten för att maximera användningen.
- Således erbjuder varje server 528 timmars körtid per månad.
- Databearbetningsuppgifter kan köra hela 528 timmar x två Dell PowerEdge XE9680-servrar = 1 056 timmars körtid.
- Tjugo dataexperter arbetar 8 timmar om dagen i 22 dagar i månaden under totalt 3 520 timmar.

Eftersom bearbetningsuppgifterna använder processorn och minnet agerar vi värd dem under de 1 056 driftstimmarna för PowerEdge XE9680-servrarna. Vi delade upp modellfinjustningen och inferenserna mellan de två servrarna med antagandet att arbetsbelastningen skulle kräva mer finjusteringstid än inferenstid. Därför beräknade vi 792 timmar per månad som läggs på finjusterande uppgifter och 264 timmar per månad på inferensuppgifter.

Slutligen antog vi att var och en av de 20 utvecklarnas användning av bärbar dator innebar en 8-timmars arbetsdag 5 dagar i veckan, totalt 3 520 timmar per månad. Hur många dataexperter ditt företag använder för att underhålla och finjustera din modell beror på flera faktorer, till exempel på hur många olika sätt du vill tolka datauppsättningen eller hur många program datauppsättningarna ska mata in. Vi valde en siffra i den högre änden av skalan för att representera en maxkostnad som kunde gälla för många företag. Eftersom dessa instanser i det offentliga molnet är mycket små och kostar lite i förhållande till lösningen som helhet, har antalet dataexperter inte en stor inverkan på den totala kostnaden för vår lösning. Med hjälp av dessa beräkningar för drifttid kunde vi mata in antalet timmar som varje instanstyp skulle köras per månad på de två molnlösningarna. Du kan se de slutliga totalkostnaderna för alla lösningar här: vetenskapen bakom rapporten.

Prisinformation för en traditionell lokal Dell-lösning

Vi kontaktade Dell och bad om en offert med rekommenderat pris för vår traditionella lokala lösning. Den här offerten inkluderade kostnaden för servrar och switchar, ProDeploy Plus för installationstjänster på plats för servrarna och en 5-årig ProSupport for Infrastructure-plan för support och underhållstjänster för utrustningen. Obs! Vi valde en 5-årig supportplan eftersom de flesta servrar håller 3 till 5 år och behöver service utöver de 3 år vi tittade på, medan vi begränsade vår totala ägandekostnad till 3 år. Vi beräknade sedan kostnaderna för energi, kylning och rackutrymme i datacentret under en period på 3 år, samt de administrativa kostnaderna för att upprätthålla utrustningen i 3 år.

Prisuppgifter för AWS SageMaker-molnlösningen

AWS delar upp sin SageMaker-tjänst i flera undertjänster som omfattar uppgifter som bearbetning och utbildning samt dataexperternas bärbara datorer. Tänk på att även om vi finjusterar en förtränad modell kallas AWS SageMaker-undertjänsten SageMaker Training. Vi använde AWS-priskalkylatorn och kalkylatorn för sparplaner för maskininlärning för att få ett pris på SageMaker-tjänsten.[5,6] I vår totala ägandekostnad prissatte vi instanser för bärbara datorer, bearbetning, finjustering av modeller och inferens enligt följande:

Tabell 2: AWS SageMaker-miljöinstanser och körtimmar per månad.

Instansmodell	Antal instanser	Uppgift	Körtid (timmar/månad)/instans
ml.t3.medium	20	Bärbar dator för dataexpert	176
ml.r5.16xlarge	2	Databehandling	1 056
ml.p5.48xlarge	1	Modellfinjustering	792
ml.p5.48xlarge	1	Inferens	264

Antaganden om prissättning:

- Vi valde två ml.r5.16xlarge-instanser för databearbetning då vi ville säkerställa minst 1 TB minne per uppgift baserat på forskning som indikerar att bearbetningsuppgifter är minnesintensiva.[7,8]
- Vi lade till 7 TB EBS-lagring per månad till ml.r5.16xlarge-instanser eftersom de inte levereras med diskar.
- Även om vi inte uppskattade kostnaderna för den lagring som är värd för den huvudsakliga datauppsättningen, uppskattade vi S3-dataöverföringskostnaderna för 1 TB in och 15 TB ut per månad för att ta hänsyn till underuppsättningarna av data som tränings- och inferensuppgifterna använder.
- Stora Ml.p5.48.-instanser var utrustade med direktansluten NVMe-lagring, och därför la vi inte till EBS-lagring för dessa instanser.

Obs! SageMaker har en elastisk strukturadapter (EFA) som erbjuder hög genomströmningshastighet.[9] Även om vi anser att nätverkshanteringen i Dell-lösningen är tillräcklig för vårt scenario kan man även välja att köpa en nätverkskonfiguration med mer bandbredd. Det är därför möjligt att AWS-lösningen kan bearbeta fler uppgifter än Dell-lösningen beroende på de nätverksval man gör.

AWS erbjuder både prissättning med användning på begäran och SageMaker-sparplaner. Prissättning med användning på begäran är den dyraste, medan sparplanerna erbjuder upp till 64 % lägre kostnader med en bindningstid på 3 år.[10] Vi prissatte AWS-konfigurationen med alternativet en bindningstid på tre år.[11] Dessutom erbjuder AWS kunderna möjlighet att betala kostnader i förskott för att få en större kostnadsminskning, något vi valde att göra för våra beräkningar av total ägandekostnad.

Jämförelse av den traditionella lokala Dell-lösningen med AWS SageMaker

Med hjälp av ovanstående antaganden för båda lösningarna beräknade vi en 3-årig jämförelse av total ägandekostnad. Våra beräkningar visar att valet av den traditionella lokala Dell PowerEdge-lösningen för att köra GenAI-arbetsbelastningar kan ge verkliga besparingar jämfört med att köra samma arbetsbelastning på AWS SageMaker.

Som bild 1 visar beräknade vi att AWS SageMaker-lösningen kunde kosta upp till 2,88 gånger så mycket som den lokala Dell-lösningen. Obs! För denna och övriga detaljer om kostnaderna kan du se vetenskapen bakom rapporten.

AWS SageMaker kan kosta nästan tre gånger så mycket som en liknande Dell-lösning

Nästan balansera utgifter och inkomster vid 1 år.

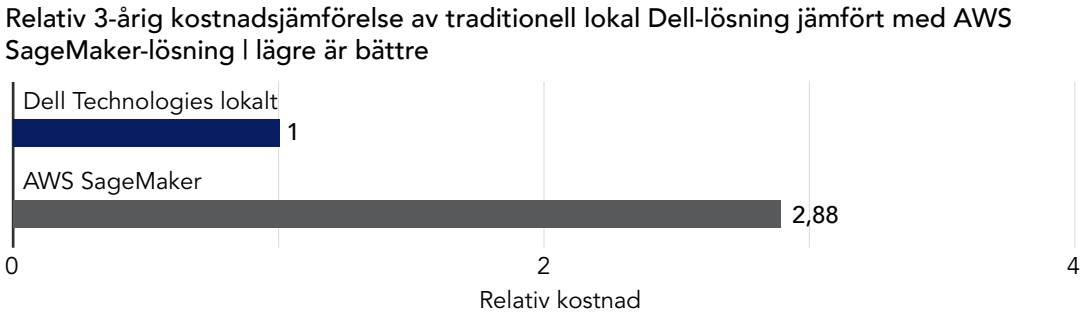

Bild 1: Relativa kostnader för en lokal GenAI Dell-lösning och en AWS SageMaker-lösning under 3 år.

Med tanke på att kostnaden under 3 år är över 2,8 gånger högre kan användarna anta att även om kostnaderna för värdtjänster, kylning och hantering av en lokal lösning ingår skulle de tjäna in kostnaden efter 1 år jämfört med kostnaden för AWS-värdtjänster.

Prisuppgifter för molnlösningen Azure Machine Learning

För Azure Machine Learning Service-miljön valde vi instanser för samma fyra uppgifter som AWS-miljön: bärbara datorer för dataexperter, databearbetning, finjustering och inferens. Vi fick vår prissättning från Azure Pricing Calculator och valde den 3-åriga planen.[12] De instanser vi har prissatt är följande:

Tabell 3: Azure Machine Learning-miljöinstanser och körtimmar per månad.

Instansmodell	Antal instanser	Uppgift	Körtid (timmar/ månad/ instans)
D2 v2	20	Bärbar dator för dataexpert	176
M64	1	Databehandling	1 056
ND96asmr A100 v4	4	Modellfinjustering	792
ND96asmr A100 v4	4	Inferens	264

Prisinformation för antaganden om Azure Machine Learning

- Azure Machine Learning-tjänsten erbjöd inte en instans med NVIDIA H100 GPU:er, så vi valde fyra A100 GPU-instanser för ungefär samma prestanda.[13]
- Alla Azure Machine Learning-instanser levereras med ansluten blocklagring, så vi prissatte inte ytterligare lagring för Azure-miljön. Precis som i våra AWS-beräkningar gjorde vi dock ungefär 10 000 000 dataöverföringsåtgärder för Azure Block Blob Storage till och från maskininlärningsinstanser.

Azure erbjuder alternativ med löpande betalning, Azure-sparplaner och Azure-reservationsalternativ för tjänsten maskininlärning.[14] Azure erbjöd inte rabatt om man betalar en engångskostnad vid inköpstillfället, som AWS gjorde. För att bäst matcha prissättningen av AWS-miljön valde vi planen med en 3-årig bindningstid.

Jämförelse av den lokala Dell-lösningen med Azure ML

Med hjälp av ovanstående antaganden beräknade vi kostnaderna för en 3-årig Azure-lösning och jämförde den med vår uppskattning av den totala ägandekostnaden under 3 år för den lokala Dell-lösningen. Återigen visar våra beräkningar att den traditionella, lokala Dell PowerEdge-lösningen för GenAI-arbetsbelastningar kan ge betydande besparingar under 3 år jämfört med en jämförbar Azure ML-lösning.

Faktum är att vi uppskattar att de totala kostnaderna för Azure Machine Learning-lösningen under 3 år blir 2,72 gånger högre än den traditionella lokala Dell-lösningen (se bild 2). Dessa resultat visar att om man lagrar hårdvara internt för GenAI med en traditionell Dell-lösning kan det bidra till att göra GenAI-budgeten ekonomiskt rimlig så att man kan använda dessa besparingar för att göra andra typer av innovationer.

Azure-maskininlärning kan kosta nästan tre gånger så mycket som en liknande Dell-lösning

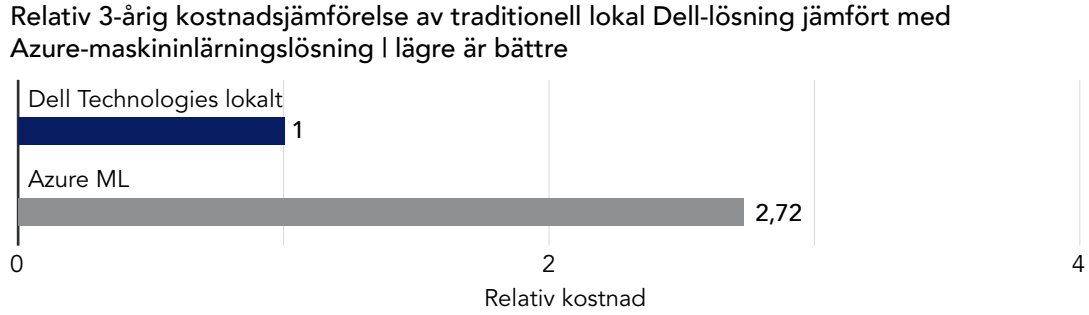

Bild 2: Relativa kostnader för en GenAI Dell-lösning på plats och en Azure Machine Learning-lösning under 3 år.

I likhet med AWS-lösningen, till över 2,7 gånger kostnaden under 3 år, kan Dells kunder med lokal lösning förvänta sig att tjäna in kostnaden inom 1 år jämfört med Azure-prissättningen.

Spara ännu mer genom att prenumerera på en Dell APEX Pay-per-Use-lösning

Vissa företag kanske upplever att det långsiktiga åtagandet i en traditionell lokal lösning är oöverkomligt. Därför Dell erbjuder en Dell APEX Pay-per-Use-lösning där man betalar efter användning. Dell kan installera hårdvara i företagets datacenter, så att den ligger lokalt som den traditionella lösningen och erbjuder en bindningstid på 3, 4 eller 5 år med en Dell APEX Pay-per-Use-lösning för beräkningsresurser vid en angiven förbrukningshastighet för en regelbunden månadsbetalning. Om du behöver mer än den avtalade förbrukningsnivån kan du utnyttja de återstående resurserna mot en extra kostnad. När prenumerationen löper ut kan du avbryta tjänsten och returnera hårdvaran, förnya i befintligt skick eller byta till en lösning som passar dina behov bättre.[15]

Inför vår jämförelse av total ägandekostnad fick vi en offert från Dell för samma hårdvara som vi inkluderade i vår traditionella lokala miljö, men vi lade också till en 3-årig prenumeration på en Dell APEX Pay-per-Use-lösning med en garanterad förbrukningshastighet på 75 %. Förbrukningshastigheterna för Dell APEX Pay-per-Use-lösning för servrar baseras på hur mycket tid en server lägger på mer än 5 % processoraktivitet under en månad. Våra antaganden var följande:

Dell APEX Pay-Per-Use-lösningar

- Ungefär 726 timmar per månad med en garanterad förbrukningshastighet på 75 % = högst 544,5 timmars servertid per månad innan ytterligare resurser behövs. Vi använde 528 timmar per månad för att garantera överensstämmelse med de andra beräkningarna.
- Offerten inkluderade även ProDeploy Plus- och ProSupport-plan för nästa arbetsdag, så vi tog inte med administratörskostnader för den första konfigurationen.
- Vi inkluderade samma kostnader för ström, kylning och rackutrymme för datacenter som vår traditionella lösning.

Vi såg att en Dell APEX Pay-per-Use-lösning som kombinerar säkerhets- och kontrollfördelarna i en traditionell lokal lösning med bekvämligheten och flexibiliteten hos en hanterad tjänst kan spara företag en betydande summa över 3 år jämfört med de molnlösningar som vi prissatt.

Som bild 3 visar kan AWS SageMaker-lösningen kosta 3,81 gånger så mycket som Dell APEX Pay-per-Use-lösningen.

AWS SageMaker kan kosta nästan 4 gånger så mycket som en liknande Dell APEX-lösning för betalning efter användning

Bli lönsam inom mindre än 1 år

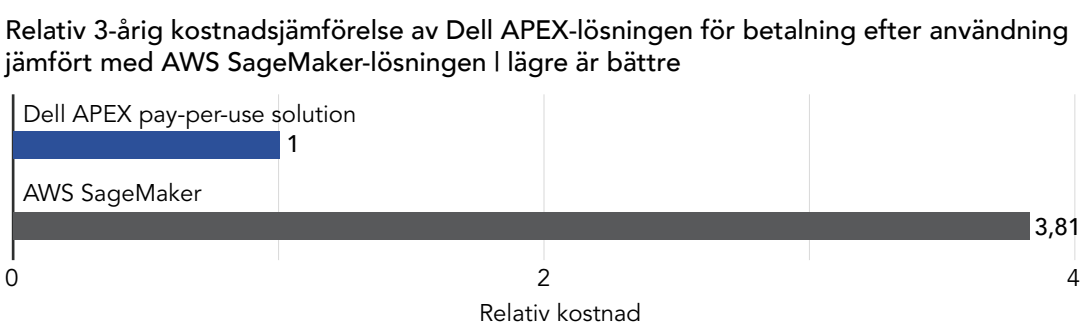

Bild 3: Relativa kostnader för en GenAI Dell APEX Pay-per-Use-lösning och en AWS SageMaker-lösning över 3 år.

Eftersom det kostar över 3,8 gånger mer än en Dell APEX Pay-per-Use-lösning under 3 år kan användarna anta att de skulle tjäna in kostnaden redan innan det första året är slut. Kostnadsbesparingarna för en Dell APEX Pay-per-Use-lösning var liknande för Azure-maskininlärningslösningen som kostar 3,60 gånger så mycket som Dell APEX Pay-per-Use-lösningen (se bild 4).

Azure-maskininlärning kan kosta nästan fyra gånger så mycket som en liknande Dell APEX-lösning för betalning efter användning

Bli lönsam inom mindre än 1 år

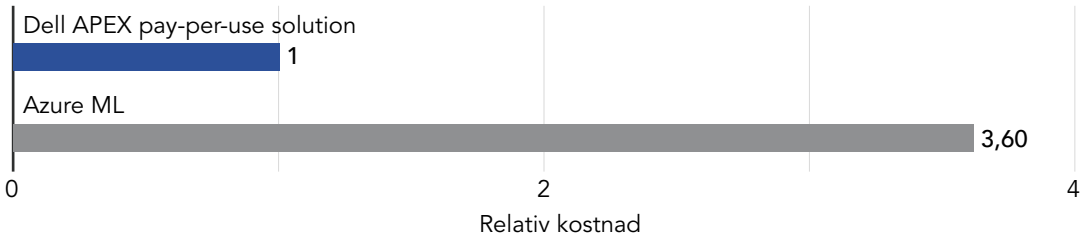

Bild 4: Relativa kostnader för en GenAI Dell APEX Pay-per-Use-lösning och en Azure-maskininlärningslösning över 3 år.

Dessa resultat visar att budgetmedvetna företag som vill implementera GenAI kan uppfylla sina behov väl genom att välja en lokal Dell APEX Pay-per-Use-lösning i stället för att vara värd för dessa potentiellt känsliga arbetsbelastningar i molnet. Förutom det kan kunder anta att de skulle tjäna in kostnaden redan efter ett år jämfört med Azure-lösningen, precis som i jämförelsen med AWS.

Ytterligare överväganden för GenAI-arbetsbelastningar

Andra fördelar med att välja en lokal lösning jämfört med molnlösning

Även om molntjänster har fördelar som skalbarhet och flexibilitet, finns det flera oroande faktorer utöver kostnad som du bör tänka på innan du blir värd för LLM:er i ett offentligt moln. En av de största är den risk som är förknippad med att lagra stora mängder användardata på en tredjepartsplattform. Många LLM:er samlar in användardata för att förbättra sina modeller och dessa data måste förvaras någonstans så att modellerna kan nå den. Lagring av känsliga data i molnet kan utsätta dem för risker som:

- Exponera data för offentliga gränssnitt som angripare kan komma åt. CrowdStrike upptäckte till exempel en sådan sårbarhet som gjorde det möjligt för dem att hitta AWS S3-buckets baserat på DNS-förfrågningar.[16]
- Ökad komplexitet kan leda till felkonfigurationer när IT-teamen ska hantera flera tjänster och molnleverantörer som regelbundet ändrar standardinställningarna.
- Risk för fel på grund av den mänskliga faktorn ökar vid användning av molnbaserade API:er som kan exponera känsliga data.[17]

Privata LLM:er minskar dessa risker eftersom användarna vanligtvis har större kontroll över sina datacenter och därmed dataströmmar, nätverksisolering, API-kontroller med mera. Dessutom har användare som kör LLM:er lokalt mer kontroll över hela stacken, från hårdvaran som LLM körs på till modellen och data som möjliggör lösningen. Administratörer kan få ytterligare utbildning för att säkerställa att lokala LLM:er följer specifika regler. I molnet har användarna mindre kontroll över den underliggande infrastrukturen och implementeringen.[18] Dessutom kan lokala lösningar göra att kostnaderna blir förutsägbara i stället för att de varierar från månad till månad.

Datalagring och överföring är en stor del av applikationskraven för LLM:er. Träning av LLM:er kräver stora mängder data som måste lagras och sedan överföras från lagring till beräkningsresurser för bearbetning. Om de enheter, databaser och användardata som ligger till grund för en LLM redan lagrar sina data lokalt kan kostnaderna bli höga för att överföra dessa data till molnet och den nätverksbandbredd som krävs.

Dell AI-portföljen

Det är tydligt att implementering av AI-arbetsbelastningar kräver mer än att bara fastställa de totala kostnaderna för lösningen. Du måste avgöra vilken modell som är bäst för dina behov och utforma en lösning som kan hantera mängden data du har. När du har valt rätt modell och bestämt dig för en lösning uppstår nästa problem: personalen. Du måste utbilda eller anställa personal för att se till att de förstår datavetenskap, kan utbilda sig i och genomföra AI-lösningen på rätt sätt.

Dell erbjuder en komplett AI-portfölj som gör mer än att ge dig kostnadsbesparingar på hårdvara: den ger de verktyg du behöver för att övervinna utmaningar under hela implementering av AI-lösningen. Dells AI-portfölj omfattar hårdvara, datahanteringstjänster, professionella tjänster, AI-utbildning, referensarkitekturer och många samarbeten med andra AI-inriktade företag.[19] Dell kan hjälpa dig att utforma, planera och implementera en framgångsrik AI-lösning som passar dina behov. Om du vill läsa mer om hur Dells AI-portfölj står sig jämfört med konkurrenternas erbjudanden kan du läsa våra rapporter där Dells AI-portfölj jämfördes med dem från Supermicro[20] och HPE.[21]

Llama 2-modeller

Llama 2, som står för Large Language Model Meta AI, är en kostnadsfri och mångsidig språkbearbetningsteknik som utvecklats av Meta. Det är en förtränad stor språkmodell (LLM) med tre huvudsakliga modellstorlekar baserade på antalet parametrar (7 B, 13 B och 70 B), som var och en ger olika prestandaegenskaper.[22] Meta tränade Llama 2 med en förstärkande inlärningsmetod för att ge familjevänligt resultat till användare, i syfte bekanta sig med mänskliga val och preferenser.[23] Läs mer om Llama 2 på https://llama.meta.com/llama2/.

Sammanfattning

Ditt företag kan få många fördelar av att dyka in i GenAI-världen. Men det kräver att du först överväger hur du bäst implementerar den arbetsbelastning som GenAI innebär. Oavsett om dina AI-mål är att skapa en chattbot för onlinebesökare, generera marknadsföringsmaterial, hjälpa till med felsökning eller något annat kräver implementering av en AI-lösning noggrann planering och beslutsfattande. Ett viktigt beslut är om du ska lagra GenAI i molnet eller behålla dina data lokalt. Traditionella lokala lösningar kan ge överlägsen säkerhet och kontroll, något som ofta är ett stort problem när man hanterar stora mängder potentiellt känsliga data. Men klarar ditt företags IT-budget att stötta en GenAI-lösning på plats?

I vår forskning upptäckte vi att värdeförslaget är precis tvärtom: att implementera en lokal GenAI-lösning, antingen som en traditionell Dell-lösning eller en hanterad Dell APEX Pay-per-Use-lösning, kan avsevärt sänka GenAI-kostnaderna under 3 år jämfört med att lagra dessa arbetsbelastningar i molnet. Faktum är att vi fann att en jämförbar AWS SageMaker-lösning skulle kosta upp till 3,8 gånger så mycket och en Azure ML-lösning skulle kosta upp till 3,6 gånger så mycket som GenAI med en Dell APEX Pay-per-Use-lösning. Dessa resultat visar att företag som vill implementera GenAI och dra nytta av de affärsfördelar som detta innebär kan hitta många fördelar med en lokal Dell-lösning, oavsett om de väljer att köpa och hantera den själva eller väljer en prenumerationsbaserad Dell APEX Pay-per-Use-lösning. Ditt företag kan spara mycket pengar genom att välja en lokal Dell-lösning jämfört med att lagra GenAI i molnet. Då får du kontroll över säkerheten och sekretessen för dina data samt eventuella uppdateringar och ändringar i miljön, och kan samtidigt säkerställa att miljön hanteras konsekvent.

---

1. LinkedIn, Jacquecine Burrells inlägg, 19 april 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”, 3 april 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”, 3 april 2024, https://oneai.com/learn/pre-trained-ai-model.
4. NVIDIA, ”What Is a Pretrained AI Model?”, 3 april 2024, https://blogs.nvidia.com/blog/what-is-a-pretrained-ai-model/.
5. AWS, ”AWS Pricing Calculator”, 3 april 2024, https://calculator.aws/#/.
6. AWS, ”Machine Learning Savings Plans”, 3 april 2024, https://aws.amazon.com/savingsplans/ml-pricing/.
7. StackOverflow, ”Why should preprocessing be done on CPU rather than GPU?”, 3 april 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu.

8. Hugging Face, ”Model Memory Requirements”, 3 april, 2024, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
9. AWS, ”Training large language models on Amazon SageMaker: Best practices”, 3 april, 2024, https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/.
10. AWS, ”Machine Learning Savings Plans”, 3 april 2024, https://aws.amazon.com/savingsplans/ml-pricing/.
11. Obs! AWS bekräftade att ml.p5.48xlarge-instansen ingår i prisplanen med en bindningstid på tre år. Vid tidpunkten för denna studie fanns den dock inte med i sparplanskalkylatorn. Vi uppskattade kostnaden för ml.p5-instansen genom att använda den besparingsprocent som anges för icke-maskininlärningsversionen p5.48xlarge enligt listan https://aws.amazon.com/savingsplans/compute-pricing/.
12. Microsoft, ”Azure Pricing Calculator”, 3 april 2024, https://azure.microsoft.com/en-us/pricing/calculator/.
13. Comet, ”Comparison of NVIDIA A100, H100 + H200 GPU:er”, 3 april 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.
14. Microsoft, ”Azure Machine Learning Pricing”, 3 april 2024, https://azure.microsoft.com/en-us/pricing/details/machine-learning/.
15. Dell, ”Dell APEX Flex On Demand”, 3 april 2024, https://www.delltechnologies.com/partner/en-us/partner/apex-flex-on-demand.htm.
16. CrowdStrike, ”12 Cloud Security issues: Risks, Threats and Challenges”, 3 april 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, ”The Pros and Cons of Using LLMs in the Cloud versus running LLMs locally”, 3 april 2024, https://www.datacamp.com/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally.
19. Dell, ”Dell AI Solutions”, 29 april 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”, 29 april 2024, https://facts.pt/q9p46K9.
21. Principled Technologies, ”Meeting the challenges of AI workloads with the Dell AI portfolio”, 29 april 2024, https://facts.pt/zPmSx4c.
22. Meta, ”Llama 2: open source, free for research and commercial use”, 3 april, 2024, https://llama.meta.com/llama2/.
23. Pavan Belagatti, ”Unpacking Metas Llama 2: The Next Leap in Generative AI”, 3 april 2024, https://www.singlestore.com/blog/a-complete-beginners-guide-to-llama2/

Vetenskapen bakom rapporten

I den här sektionen listar vi alla våra resultat och beskriver lösningarna som vi testade och våra testmetoder.

Vi slutförde vår forskning den 29 mars 2024. Resultaten i den här rapporten återspeglar konfigurationer som vi slutförde och förvärvade prisdata för den 29 mars 2024 eller tidigare. Det innebär att dessa konfigurationer och deras kostnader kanske inte är uppdaterade när den här rapporten visas.

Systeminformation

Dells lokala lösningar

De traditionella och hanterade lokala Dell-lösningarna omfattar följande hårdvara:

- 3 x PowerEdge R660-huvudnod
- 2 x PowerEdge XE9680 GPU-arbetsnod
- 2 x PowerSwitch S5232-ON-nätverksinfrastruktur
- 1 x PowerSwitch N3200-ON OOB-hantering

Tabell 1: : Detaljerad konfigurationsinformation för varje PowerEdge XE9680 GPU-arbetsnod.

Konfigurationsinformation	Dell PowerEdge XE9680 GPU-arbetsnod
Antal noder i lösningen	2
Chassi	
Chassit	XE9680 6U-chassi med endast 8 GPU 8x 2,5 NVMe
Processor	
Antal processorer	2
Leverantör och modell	Intel® Xeon® Platinum 8468
Antal kärnor (per processor)	48 kärnor och 96 trådar
GPU:er	
Antal GPU:er	1 8-GPU-montering
Leverantör och modell	NVIDIA® HGX H100 8-GPU SXM 80 GB 700 W GPU-montering
Minnesmodul(er)	
Totalt minne i systemet (GB)	1 024
Antal minnesmoduler	16
Typ	RDIMM, 4 800 MT/s Dual Rank
Storlek (GB)	64
Lagringsstyrenhet	
Leverantör och modell	BOSS-N1-styrkort + med två M.2 480 GB (RAID 1)
Lokal lagring (typ A)	
Total storlek på enheter i systemet (TB)	44.8
Antal diskar	7

Konfigurationsinformation	Dell PowerEdge XE9680 GPU-arbetsnod
Drivenhetsstorlek (TB)	6.4
Drivenhetsinformation (hastighet, gränssnitt, typ)	Enterprise NVMe™ för blandad användning AG Drive U.2 Gen4 med hållare
NIC	
Antal och typ av portar	2 x 10/25GbE
Leverantör och modell	Intel E810-XXV Dual Port 10/25 GbE SFP28, OCP NIC 3.0
Nätverksadapter	
Antal och typ av portar	2 x 10/100GbE
Leverantör och modell	Mellanox ConnectX-6 DX med två portar 100 GbE QSFP56 nätverksadapter, full höjd
Kylfläktar	
Antal kylfläktar	6
Leverantör och modell	Fläkt med mycket hög prestanda
Nätaggregat	
Antal nätaggregat	1
Leverantör och modell	Helt redundant 5 + 1 (eller 3 + 3 FTR), nätaggregat som kan bytas under drift, 2 800 W MM HLAC (200–240 V AC) Titanium, C22-kontakt
Effekt per enhet (W)	2800
ProSupport och ProDeploy	
ProSupport (5 år)	ProSupport och service på plats nästa arbetsdag
ProDeploy Plus	ProDeploy Plus Dell-server i XE-serien 5U/6U
Inbyggd systemhantering	
iDRAC9	iDRAC9, Datacenter 16 G
OpenManage	OpenManage™ Enterprise Advanced Plus

Tabell 2: Detaljerad konfigurationsinformation för varje PowerEdge R660-huvudnod.

Konfigurationsinformation	Dell PowerEdge R660 huvudnod
Antal noder i lösningen	3
Chassi	
Chassit	2,5-tums chassi med upp till 10 hårddiskar (SAS/SATA), PERC11, 1 processor
Processor	
Antal processorer	1
Leverantör och modell	Intel Xeon Gold 6426Y
Antal kärnor (per processor)	16 kärnor och 32 trådar
Minnesmodul(er)	
Totalt minne i systemet (GB)	192
Antal minnesmoduler	12

Konfigurationsinformation	Dell PowerEdge R660 huvudnod
Typ	16
Storlek (GB)	RDIMM, 4 800 MT/s Single Rank
Lagringsstyrenhet	
Leverantör och modell	BOSS-N1-styrkort + med 2 SED M.2 960 GB (RAID 1)
Lokal lagring (typ A)	
Total storlek på enheter i systemet (TB)	5.76
Antal diskar	6
Drivenhetsstorlek (TB)	960
Drivenhetsinformation (hastighet, gränssnitt, typ)	vSAS Read Intensive SSD 12 Gbit/s 512e 2,5 tum, kan bytas under drift, AG Drive SED, 1DWPD
NIC	
Antal och typ av portar	2 x 10/25GbE
Leverantör och modell	NVIDIA ConnectX-6 Lx med dubbla portar 10/25 GbE SFP28, ingen krypto, OCP NIC 3.0
Kylfläktar	
Antal kylfläktar	4
Leverantör och modell	Fläkt med mycket hög prestanda
Nätaggregat	
Antal nätaggregat	1
Leverantör och modell	Dubbel, redundans(1+1), nätaggregat som kan bytas under drift, 11 000 W MM(100-240 VAC) Titanium
Effekt per enhet (W)	1 100
ProSupport och ProDeploy	
ProSupport (5 år)	ProSupport och service på plats nästa arbetsdag
ProDeploy Plus	ProDeploy Plus PowerEdge R-serien 1u/2u
Inbyggd systemhantering	
iDRAC9	iDRAC9, Datacenter 16 G
OpenManage	OpenManage Enterprise Advanced Plus

AWS SageMaker-lösningsinstanser

Tabell 3: Detaljerad konfigurationsinformation för AWS-instanser.

Konfigurationsinformation	ml.t3.medium (bärbara datorer)	ml.r5.16xlarge (bearbetning)	ml.p5.48xlarge (inferens och finjustering)
Antal instanser	20	2	2
Molntjänstleverantörer (CSP)	AWS	AWS	AWS
Region	USA, östra (Ohio)	USA, östra (Ohio)	USA, östra (Ohio)
Processor			
Antal vCPU	2	64	192

Konfigurationsinformation	ml.t3.medium (bärbara datorer)	ml.r5.16xlarge (bearbetning)	ml.p5.48xlarge (inferens och finjustering)
Minnesmodul(er)			
Totalt minne i systemet (GB)	4	512	2 048
Lagringsstyrenhet			
Leverantör och modell			
Lokal lagring (typ A)			
Antal diskar	1	1	1
Drivenhetsstorlek (GB)	5 GB	Standard	Standard
Drivenhetsinformation (hastighet, gränssnitt, typ)	EBS	EBS	EBS
Grafikprocessor			
Antal GPU:er	e.t.	e.t.	8
Leverantör och modell	e.t.	e.t.	NVIDIA H100
Fler funktioner	e.t.	e.t.	3 200 Gbit/s nätverksbandbredd[1]

Azure-lösningsinstanser för maskininlärning

Tabell 4: Detaljerad konfigurationsinformation för Azure-instanser.

Konfigurationsinformation	D2 v2 (bärbara datorer)	M64 (bearbetning)	ND96amsr A100 v4 (inferens och finjustering)
Antal instanser	20	1	8
Molntjänstleverantörer (CSP)	Azure	Azure	Azure
Region	Östra USA 2	Östra USA 2	Östra USA 2
Processor			
Antal vCPU	2	64	96
Minnesmodul(er)			
Totalt minne i systemet (GB)	7	1 000	1 900
Lokal lagring (typ A)			
Antal diskar	1	1	1
Drivenhetsstorlek (GB)	100	7 168	6 400
Drivenhetsinformation (hastighet, gränssnitt, typ)	Temporär	Temporär	Temporär
Antal GPU:er	e.t.	e.t.	8
Leverantör och modell	e.t.	e.t.	NVIDIA A100
Ytterligare information	e.t.	Ej tillämpligt	Varje GPU har en dedikerad 200 GB/s NVIDIA Mellanox HDR InfiniBand-anslutning[2]

Inledning

Vi skapade ett AI-scenario med hjälp av Llama 2 13B-modellen med öppen källkod för att ge en uppfattning om hur mycket AI-lösningar kostar. Här jämförde vi kostnaden för att köra arbetsbelastningen i fyra olika miljöer. Vi dimensionerade och beräknade kostnaderna för fyra lösningar:

- Traditionell och lokal Dell-lösning
- Hanterad lokal lösning med en Dell APEX Pay-per-Use-lösning
- AWS SageMaker-lösning
- Microsoft Azure Machine Learning-lösning

Både traditionella och hanterade lokala Dell-lösningar använder samma hårdvara. Med den traditionella planen betalar företaget hårdvaran i förskott. Med Dell APEX Pay-per-Use-lösning installerar Dell hårdvara i kundens datacenter och fakturerar företaget månadsvis baserat på ”avtal” och ”buffertkapacitet”.

För den här analysen försökte vi skapa ett brett tillämpligt exempelscenario för att uppskatta kostnadsskillnader i olika miljöer. Vi valde Llama 2 13B GenAI-modellen eftersom den är en allmänt tillgänglig modell med öppen källkod, och vi byggde vårt scenario kring en enda och relativt liten AI-arbetsbelastning. Vi inkluderade kostnader för dataexperternas bärbara datorer som används för utveckling av maskininlärning, databearbetningsuppgifter, kontinuerlig modellfinjustering och realtidsinferens.

Vi dimensionerade hårdvaran för varje lösning baserat på antaganden om antalet arbetstimmar och vilken hårdvarukapacitet som behövs. Vi använde offentliga källor för den forskningen. Vi använde priskalkylatorer online för AWS SageMaker och Azure Machine Learning. Vi begärde och fick offerter från Dell Technologies för kostnaderna med hjälp av Dells rekommenderade priser för de två Dell-lösningarna. Vi utförde inga praktiska tester av några av de lösningar som användes för denna rapport.

Våra resultat

I huvudrapporten visar vi normaliserade jämförelser för var och en av de två lokala Dell-lösningarna jämfört med molnlösningarna AWS SageMaker och Azure Machine Learning. I tabell 5 och 6 visas kostnadsbasen för dessa normaliseringar. Det normaliserade värdet är resultatet av att varje värde divideras med kostnaden för Dells lokala lösning som visas i tabellen. Tabell 5 visar att molnlösningar kostar upp till 2,88 gånger Dells traditionella lokala lösning under tre år. Raden för att tjäna in kostnaden visar att tre år av den traditionella lokala Dell-lösningen kostar ungefär lika mycket som kostnaden för bara ett år av de två molnlösningarna.

Tabell 5: Normaliserad och treårig total ägandekostnad för Dells traditionella lokala lösning jämfört med SageMaker- och Azure Machine Learning-lösningar.

	Kostnader för Dells traditionella lokala lösning under 3 år	AWS SageMaker med bindningstid på 3 år	Azure Machine Learning med bindningstid på 3 år
Totalt	817 880,00 USD	2 357 549,00 USD	2 231 805,00 USD
Normal	1	2,88	2,72
Breakeven i månader för Dells lösning jämfört med molnlösningar	e.t.	12.5	13,3

Tabell 6 visar att molnlösningarna uppgår till upp till 3,81 gånger kostnaden för en Dell APEX Pay-per-Use-lösning över 3 år och att 3 år med Dells traditionella lokala lösning kostar mindre än kostnaden för bara 1 år med de två molnlösningarna.

Tabell 6: Normaliserad 3-årig total ägandekostnad för en Dell APEX Pay-per-Use-lösning jämfört med SageMaker- och Azure Machine Learning-lösningarna.

	Dell APEX Pay-per-Use-lösning, kostnad under 3 år	AWS SageMaker med bindningstid på 3 år	Azure Machine Learning med bindningstid på 3 år
Totalt	618 648,00 USD	2 357 549,00 USD	2 231 805,00 USD
Normal	1	3,81	3,60
Breakeven i månader för Dells lösning jämfört med molnlösningar	e.t.	9.5	10

Lagringsöverväganden

Vi tog inte med kostnader för lagring utöver vad servrarna eller instanserna behöver för att utföra sina uppgifter.

Dells lokala lösningar inkluderar 106,88 TB lagring:

- 5,76 TB SSD-kapacitet på var och en av de tre PowerEdge R660-huvudnoderna
- 44,8 TB SSD-kapacitet på var och en av de två PowerEdge XE9680 GPU-arbetsnoderna

Klusterhantering och uppgifter för bärbara datorer delar lagringsutrymmet på huvudnoderna. Behandlings-, modellfinjusterings- och inferensarbetsuppgifter delar lagringsutrymmet på GPU-arbetsnoderna. Vi utrustade Dell PowerEdge-klustren med lite extra lagringsutrymme jämfört med molnlösningarna för att säkra utrymme för hantering av uppgifter på huvudnoderna och lite utrymme för tillväxt om det skulle behövas.

AWS SageMaker-lösningen inkluderar totalt 63 444 TB lagringsutrymme:

- 7 000 GB EBS gp2-lagring som köpts för var och en av de två ml.r5.16xlarge-bearbetningsinstanserna
- 8 x 3 084 GB NVMe SSD-diskar för var och en av ml.p5.48xlarge-instanserna
- 1 x 5 GB EBS tillfällig lagring för varje bärbar instans som ingår i instansen

Azure Machine Learning-lösningen inkluderade totalt 64,82 TB temporärt lagringsutrymme:

- 7 168 GiB tillfällig lagring för M64-bearbetningsinstansen
- 6 400 GiB tillfällig lagring för var och en av de åtta ND96amsr A100 v4-instanserna
- 1 x 100 GiB tillfällig lagring för varje bärbar instans som ingår i instansen

Lagringsbehoven varierar för instanser för bärbara datorer, men kräver vanligtvis inte mycket för den här typen av arbetsbelastning. Därför valde vi att låta molninstanser behålla den lagring som ingår som standard. Vi inkluderade kostnader för dataöverföring för EBS-dataöverföring i AWS SageMaker och för Blob-lagringsöverföring i Azure Machine Learning-lösningar. Vi tog inte med kostnader för dataöverföring för Dells lokala lösningar som använder inbyggda SSD-diskar.

Användningstimmar

Vi dimensionerade lösningarna baserat på följande uppskattningar av timmar per månad för bärbar dator, databehandling, modellfinjustering och inferensanvändning. Vi använde dessa timmar för att beräkna instansanvändning för molnlösningarna och för att dimensionera dessa instanser och servrarna för lokala lösningar.

Vi dimensionerade lösningarna med antagandet att det är 22 arbetsdagar i varje månad, med arbetsbelastningar som är inställda på att köras över natten för att maximera användningen. Varje server- och molninstans skulle alltså ha 528 timmars drifttid tillgänglig varje månad. (Se tabell 7.)

Tabell 7: Användningstimmar för de fyra uppgifterna.

Uppgift	Totalt antal timmar per månad	Beräkningar av användning
Bärbar dator	3 520	Vi dimensionerade varje lösning för att stötta 20 dataexperter med en bärbar datorinstans var, 8 timmar om dagen, 22 dagar i månaden. Totalt 176 timmar varje månad och totalt 3 520 för alla 20 dataexperter. Minimikrav är små bärbara molninstanser med 2 vCPU och minst 4 GiB-minne.
Bearbetar	1 056	Databearbetningsuppgifter skulle köras under 528 drifttimmar på två Dell PowerEdge XE9680-servrar för totalt 1 056 timmars körtid och skulle kräva 1 056 timmars körtid på molninstanser. Minimikrav för molninstanser var 64 vCPU, 1 000 GiB-minne, 7 000 GB lagring. Vi dimensionerade Dell PowerEdge-servrarna för att stötta dessa krav plus kraven för finjustering och slutledning.
Modellfinjustering	792	Den kombinerade körtiden på 1 056 timmar för finjustering och slutledning kräver två Dell PowerEdge XE9680-servrar. Alla jobb använder åtta H100-grafikprocessorer och upp till en halv TB minne. Molnlösningarna kräver samma antal timmar på instanser med åtta H100-GPU:er eller motsvarande. För AWS använde vi två H100-instanser; för Azure Machine Learning bytte vi ut mot åtta A100-instanser.
Inferens	264	

Information om bärbara datorer

Dataexperter skulle behöva små bärbara molninstanser med 2 vCPU och minst 4 GiB-minne. Även om vissa uppgifter i bärbara datorer kanske fungerar bättre med mer minne valde vi AWS ml.t3. medelstor instans baserat på förslagen i AWS SageMaker TCO Guide[3] och valde sedan en liknande storleksinstans för Azure. För Dells lokala lösningar antog vi en kärna i en motsvarande processor och 4,3 GB minne per bärbar dator som körs på 3 PowerEdge R660-hanteringsservrarna tillsammans med hanteringsuppgifter.

Instanser för AWS SageMaker och Azure Machine Learning

För AWS SageMaker och Azure Machine Learning valde vi instanser för bärbara datorer som hade 2 vCPU och minst 4 GiB-minne.

Tabell 8: Viktig konfigurationsinformation för instanser för AWS SageMaker- och Azure Machine Learning.

Instanstyp	Instans	Antal vCPU per instans	Minneskapacitet (GiB) per instans
SageMaker bärbara datorer	ml.t3.medium	2	4
Azure ML bärbara datorer	D2 v2	2	7

Dells lokala lösningar för bärbara datorer

Dells lösningar stöder dessa arbetsbelastningar för bärbara datorer på de tre PowerEdge R660-huvudnoderna. Tillsammans har de 576 GB minne och 48 processorkärnor, tillräckligt för att stötta både klusterhanteringsuppgifter och arbetsbelastningar för 20 bärbara datorer. Våra antaganden när vi fattade beslutet om dimensionering är följande:

- Hanteringsuppgifter tar upp mindre än hälften av processorkapaciteten hos dessa huvudnoder och mindre än 500 GB minne, med återstående kapacitet tillgänglig för att köra dessa uppgifter.
- Under de 176 timmar som varje arbetsbelastning för bärbara datorer kör varje månad skulle den använda en enda processorkärna, motsvarande 2 vCPU för de bärbara SageMaker- och Azure ML-datorerna, med en tråd: 1 vCPU-förhållande och 4,3 GB minne som matchar de 4 GiB vi definierade i dimensioneringen av de bärbara datorerna. Alla 20 bärbara datorer som körs samtidigt skulle använda mindre än hälften av de 48 kärnorna och 15 % av minnet på dessa system.
- Allt arbete på alla system sker under mindre än 72,3 % av den totala tiden varje månad, baserat på 22 arbetsdagar i veckan med 24 timmar tillgängliga varje dag.

Bearbetningsdetaljer

Bearbetningsinstanser för AWS SageMaker och Azure Machine Learning

Bearbetningsuppgifter körs bäst på processorn i stället för på grafikprocessorn och lyckas bäst med ett högt förhållande mellan minne och kärna[4], så vi fokuserade på minnesoptimerade instanser för AWS SageMaker och Azure Machine Learning. Vårt mål var minst 64 vCPU, 1 000 GiB-minne och 7 000 GB lagring, något som bearbetningsinstansen för Azure Machine Learning M64 nästan matchade med 64 vCPU, 1 000 GiB-minne och 7 168 GiB tillfälligt diskutrymme. AWS SageMakers minnesoptimerade bearbetningsinstanser hade inte någon instans som matchar vår specifikation. Istället valde vi ett par ml.r5.16xlarge SageMaker minnesoptimerade bearbetningsinstanser med ett kombinerat minne på 1 024 GiB. Tillsammans överstiger dessa SageMaker-instanser våra behov med dubbel vCPU-kapacitet och nästan dubbelt så stor lagringskapacitet (med EBS-lagring) som våra mål och Azure Machine Learning-instansen.

Tabell 9: Viktig konfigurationsinformation för bearbetningsinstanser för AWS SageMaker och Azure Machine Learning.

Information om instanskonfiguration	SageMaker ml.r5.16xlarge (bearbetning)	Azure Machine Learning M64 (bearbetning)
Antal instanser	2	1
Molntjänstleverantörer (CSP)	AWS	Azure
Antal vCPU	64 (128 för 2 instanser)	64
Totalt minne i systemet (GB)	512 (1 024 för två instanser)	1 000
Antal diskar	1 (2 för två instanser)	1
Enhetsstorlek (GiB)	7 000 (14 000 för två instanser)	7 168
Drivenhetsinformation (hastighet, gränssnitt, typ)	EBS	Temporär

Dells lokala lösningar

De två Dell PowerEdge XE9680-arbetsnoderna har mer kapacitet än de finjusterings- och inferensarbetsbelastningar som krävs, och är tillräckligt för att stötta arbetsbelastningar som körs parallellt. Arbetsbelastningarna är beroende av processorer, finjustering av modell och inferens på GPU:er. För att matcha våra målspecifikationer skulle arbetsbelastningarna använda hälften av det kombinerade minnet på 2 TB hos de två servrarna, 32 processorkärnor och cirka 7 TB av de två servrarnas lagringskapacitet på 44,8 TB.

Finjustering av modell och inferensdetaljer

Lösningarna kräver GPU-instanser eller servrar för finjustering och inferensarbetsbelastningar med åtta NVIDIA HGX H100-GPU:er eller motsvarande GPU-kapacitet och 512 GB minne.

Instanser för AWS SageMaker and Azure Machine Learning

Ml.p5.48xLarge-inferensinstansen är den enda SageMaker-instansen som har NVIDIA HGX H100 GPU:er.[5] Vid tidpunkten för vår studie hade den bästa Azure VM åtta NVIDIA A100 GPU VM:er, så vi ställde in Azure-miljön att köra fyra gånger så många instanstimmar för att uppnå ungefär samma prestanda som H100s i AWS SageMaker-miljön.[6] Båda instanstyperna har mer än vårt totala minneskrav på 512 GB.

Tabell 10: : Viktig konfigurationsinformation för finjustering och inferensinstanser för AWS SageMaker och Azure Machine Learning.

Inferens och utbildningsinstanser	SageMaker ml.p5.48xlarge	Azure Machine Learning ND96amsr A100 v4
Antal instanser	2 (1 för finjustering och 1 för inferens)	8 (4 för finjustering och 4 för inferens)
Molntjänstleverantörer (CSP)	AWS	Azure
Antal vCPU	192	96
Totalt minne i systemet (GB)	2 048	1 900
Antal diskar	8	1
Enhetsstorlek (GiB)	3 084	6 400
Drivenhetsinformation (hastighet, gränssnitt, typ)	NVMe SSD	Temporär
Antal GPU:er	8	8
Leverantör och modell	NVIDIA H100	NVIDIA A100

Dells lokala lösningar

För Dells lokala lösningar dimensionerade vi de två PowerEdge XE9680 GPU-arbetsnoderna för att hantera finjusterings- och inferensarbetsbelastningar med hjälp av GPU-resurser och för att stödja bearbetningsarbetsbelastningar med hjälp av extra CPU- och minnesresurser. De två GPU-arbetsnoderna för PowerEdge XE9680 har vardera två 48-kärniga Intel Xeon Platinum 8468-processorer, 1 024 GB minne, 44,8 TB lagring och en NVIDIA HGX H100 8-GPU-enhet.

Kostnadsanalys

För molnlösningarna inkluderade vi licenskostnaden för de instanser vi behövde för bärbara datorer, bearbetning, finjustering och inferens. För Dells lokala lösningar inkluderar vi två betalningsalternativ för hårdvaran: en modell för direktköp och en månatlig betalningsplan för Dell APEX Pay-per-Use-lösning. För båda de lokala lösningarna la vi till kostnader för serveradministration för hårdvara och operativsystem, samt kostnader för datacenter för rackutrymme och energikostnader för ström och kylning. Dessa kostnader är inte relevanta för de två molnlösningarna.

Vi utelämnade vissa kostnader, till exempel:

- Vi utelämnade kostnader för arbete som skulle kunna se likadant ut för alla fyra lösningar: installera och underhålla programvara med öppen källkod, dataöverföring och säkerhetskopiering samt löner för dataexperterna.
- Vi inkluderade inte programvarukostnader för någon av lösningarna. Instanser för SageMaker och Azure Machine Learning innehåller vissa program och tjänster, till exempel Jupyter Notebooks på de bärbara datorerna och bearbetnings-API:er med bearbetningsinstanserna. Vi antog att ytterligare programvara och verktyg som dataexperterna installerar där skulle vara öppen källkod. Med Dells lokala lösningar skulle dataforskare enbart använda programvara och verktyg med öppen källkod som Jupyter Notebook, Python och PyTorch, och servrarna skulle köra Ubuntu eller ett annat operativsystem med öppen källkod.

- Vi inkluderade inte försäljningsskatter eftersom de varierar för olika stater och företag. Vi inkluderar inte migreringskostnader eller kostnader för utrustning som närmar sig slutet på livslängden.

Vi fokuserade på en 3-årig livscykel för var och en av dessa generativa AI-lösningar. Vi valde 3 år eftersom AWS och Azure Pricing Calculator begränsade bindningstiden till tre år. Tre år är också en rimlig livscykel för en lokal generativ AI-lösning som kräver toppmodern hårdvara.[7] Företag skulle kunna återanvända den Dell-hårdvara de köpte efter det och skulle då ha två återstående år av de fem år som ingår i Dell ProSupport-tjänsterna för att hjälpa till att hålla IT-produktiviteten uppe.

Kostnader för Dells lokala lösningar under en 3-årsperiod

För Dells lokala lösningar inkluderade vi följande kostnader under en 3-årsperiod:

- Dells rekommenderade pris för Dells servrar och switchar, inklusive ProSupport och service på plats nästa dag och ProDeploy Plus för servrarna
- Systemadministratör för att underhålla och skydda hårdvara och operativsystem
- Energikostnader för strömförsörjning och kylning
- Kostnader för rackutrymme i datacenter

De två lösningarna inkluderade samma hårdvara och skulle medföra samma kostnader för systemadministration, energikostnader för strömförsörjning och kylning samt kostnader för rackutrymme i datacenter.

Tabell 11: Kostnader för traditionella lokala lösningar under en 3-årsperiod.

Dells traditionella lokala lösning	Kostnader under tre år (avrundat upp till dollar)
Dells hårdvara med 5 års ProSupport och ProDeploy Plus (för servrar)	680 251 USD
Systemadministration	6 531 USD
Energikostnader för strömförsörjning och kylning	88 978 USD
Kostnader för rackutrymme i datacenter	42 120 USD
Totalt	817 880 USD

Vi begärde en offert från Dell Technologies Sales för traditionella lokala lösningar och det rekommenderade priset för den lokala Dell-lösningen. Den 20 mars 2024 gav Dell oss en offert på 680 251 USD som inköpspris för hårdvaran.

Offerten inkluderade 5 års ProSupport och service på plats nästa arbetsdag för servrar och switchar och ProDeploy Plus för servrarna. Dell definierar rekommenderat pris som ett pris som fungerar som en utgångspunkt för potentiella köpare. Det motsvarar den kostnad som omedelbart är tillgänglig för företag, även om de inte är befintliga kunder, och fungerar huvudsakligen som ett rekommenderat detaljhandelspris för produkterna.

Tabell 12: Kostnader under 3 år för en Dell APEX Pay-per-Use-lösning.

Dell APEX Pay-Per-Use-lösning	Kostnader under tre år (avrundat upp till dollar)
Dells hårdvara med 5 års ProSupport och ProDeploy Plus (för servrar)	481 019 USD
Systemadministration	6 531 USD
Energikostnader för strömförsörjning och kylning	88 978 USD
Kostnad för golvyta i datacenter	42 120 USD
Totalt	618 648 USD

Lokal Dell APEX Pay-per-Use-lösning

Baserat på det rekommenderade priset gav Dell Technologies en kostnadsuppskattning på 3 år för en Dell APEX Pay-per-Use-lösning på 481 019 USD. Dell skickade PT en offert för Dell APEX Pay-per-Use-lösningen för samma hårdvara som ovan med en bindningstid på 36 månader och ett kapacitetsåtagande på 75 %. Vi fick den offerten den 2 april 2024. Den uppskattade kapaciteten på 75 % var det närmaste kapacitetsalternativet som skulle täcka de 528 drifttimmar som vi uppskattar att lösningarna kommer att leverera.

Systemadministration

Serveradministratörer övervakar och säkerställer prestanda, tillgänglighet, funktion och säkerhet för maskinvaran och operativsystemet, och i det här fallet installation av operativsystemet. Det här är tjänster som den lokala lösningen kräver, men molnlösningarna gör det inte då detta ingår i serviceavtalet. Vi höll dessa tids- och kostnadsberäkningar låga för den lokala lösningen eftersom vi antar att de mestadels är automatiserade och eftersom ProSupport for Infrastructure avlastar vissa av övervakningsuppgifterna och de fysiska reparationsuppgifterna från de lokala systemadministratörerna.

Vi uppskattade en treårig kostnad på 6 531 00 USD för den här serveradministrationen baserat på den totala ersättningen för en systemadministratör[8] som kan underhålla 300 servrar och tillhörande switchar och operativsystem med hjälp av automatiserade verktyg och processer och som får hjälp av ProSupport- och ProDeploy Plus-tjänster.

Dell ProDeploy Plus for Infrastructure

Vi tog inte med en separat uppskattning för distribution, utan förlitar oss istället på Dell ProDeploy Plus for Infrastructure, en tjänst som vi inkluderade i hårdvaruofferten för att tillhandahålla hårdvaru- och mjukvarudistribution på plats.[9] En rapport från Principled Technologies visar att ProDeploy Plus for Infrastructure kan ”spara värdefull intern administratörstid genom att använda en Dell Technologies-certifierad tekniker för installation och konfiguration av en Dell”.[10]

Den tjänst som erbjuds kan inte täcka alla planeringsuppgifter, att packa upp och montera switchar, eller att installera operativsystemet. Dessa uppgifter skulle ta lite tid, och ingår i uppskattningen av systemadministrationstiden för att underhålla och säkra lösningen.

Dell ProSupport for Infrastructure

Vi inkluderade 5 års Dell ProSupport och service på plats nästa dag. Den 5-åriga supporten är längre än den 3-åriga tidsperioden för vår analys, men ger företaget som köper tjänsten det mervärde som en längre livscykel ger för Dell-hårdvaran.

Energikostnader för strömförsörjning och kylning

Molnlösningarna inkluderade energikostnader för ström och kylning i sina priser. För de lokala lösningarna uppskattade vi tre års ström- och kylningskostnader med hjälp av verktyget Dell Enterprise Infrastructure Planning.[11] För att få en uppskattning har vi angett specifikationerna för de servrar och switchar som ingår i Dell Technologies försäljningsprisoffert. Vi angav två andra indata som påverkade beräkningarna:

- Energieffektivitet (PUE) med multipliceringsfaktorn 1,58 för att få kombinerade kostnader för ström och kylning. Att energieffektiviteten var branschgenomsnittet 2023, enligt Uptime Institute, en organisation som granskar och spårar datacenterkostnader.[12]
- 12,74 öre per kilowattimme i energikostnad baserat på amerikanska EIA:s rapporterade genomsnittliga detaljhandelspris på el för den kommersiella sektorn under 2023.[13]

Vi beräknade kostnader för ström och kylning separat för inaktiva enheter och beräkningsarbetsbelastningar. Vi viktade resultaten baserat på de 528 drifttimmar (cirka 72,3 % av en genomsnittlig månad) som vi dimensionerade lösningarna för att leverera.

Tabell 13: Energikostnad för ström och kylning under en 3-årsperiod.[14]

Arbetsbelastningar	Energikostnad för lokal lösning under en 3-årsperiod	Viktad multipliceringsfaktor	Viktad energikostnad för ström och kylning
Beräkning	114 439,69 USD	72,3%	82 739,90 USD
Inaktiv	22 516,69 USD	27,7%	6 237,12 USD
Viktad energikostnad för ström och kylning under en 3-årsperiod			88 977,02 USD

Kostnad för rack i datacenter

Företaget skulle få extra kostnader för att förvara servrarna och racken i datacentret. Dessa kostnader inkluderar kostnaderna för rack, nätverk och andra datacenteranläggningar. Vi uppskattade dessa kostnader till 1 800 USD per rack och månad för rack med en användbar kapacitet på 28U.[15] servrarna och switcharna i den här lösningen fyller 18u, 65 % av ett av dessa rack. Kostnaden över tre år för denna användning är 42 120 USD.

Även om vi inkluderade dataöverföring för de två molnlösningarna för åtkomst till AWS S3-lagring och Azure-blocklagring, la vi inte till dessa kostnader för de lokala lösningarna eftersom de skulle använda sina inbyggda diskar eller lokala lagringsdisksystem för lagring.

AWS SageMaker-lösning

Vi konfigurerade AWS SageMaker-lösningen så att den matchar den lokala Dell-lösningen så nära som möjligt för de fyra uppgifter vi beskrev tidigare: bärbara datorer, bearbetning, modellfinjustering och inferens. AWS-priskalkylatorn för SageMaker listar varje uppgift som en separat prismodul som du kan växla för att lägga till i beräkningen. Vi lade till SageMaker Studio Notebooks, SageMaker Processing, SageMaker Training och SageMaker Real-Time Inference. För varje modul fyller vi i de nödvändiga fälten för att bestämma timkostnaden för varje instans vi valde för varje uppgift. Vi använde sedan den timkostnaden för att avgöra hur mycket en användare skulle spendera för att köra varje instans för det förberäknade antalet timmar vi tog fram baserat på Dell-systemen. Vi beräknade dubbelt så många bearbetningsinstanser och därmed bearbetningstimmar för att säkerställa att de matchar bearbetningskapaciteten för de andra lösningarna. Vi har även lagt till EBS-lagring i bearbetningsinstanserna, eftersom de inte har lagring utanför OS-volymen. Efter att ha tillämpat hela den initiala 3-åriga besparingsplanen uppgick våra totala kostnader till 2 357 549 USD under 3 år. Fullständig information om instansen finns i tabell 14.

Tabell 14: Instanskostnad för SageMaker-lösning under en 3-årsperiod.

Tjänst	Instanstyp	Instans, USD/tim	Körtid (timmar/månad)	Kostnad för 3 år
SageMaker Studio bärbara datorer	ml.t3.medium	0,02244 USD	3520	2 843,60 USD
SageMaker-bearbetning	ml.r5.16xlarge*	2 2908 USD	2112	174 174,11 USD
SageMaker-bearbetning av EBS-lagring (7 TB per månad)				25 804,80 USD
SageMaker-utbildning	ml.p5.48xlarge**	56 5340 USD	792	1 611 897,41 USD
SageMaker-inferens i realtid	ml.p5.48xlarge**	56 5340 USD	264	537 299,14 USD
S3-dataöverföring (1 in och 15 ut)				5 529,60 USD
Totalt (avrundat uppåt)				2 357 549 USD

*Vi inkluderar processorinstanser för att få 1 TB minne för bearbetningsuppgifter.
**Priser som uppskattas baserat på offentlig P5-avtalsprocent (ej ml.p5) eftersom kalkylatorn inte inkluderade den här instansen.

Azure Machine Learning-lösning

Vi använde Azure Pricing Calculator för att ange varje instanstyp och fastställa timkostnaden för varje typ i Machine Learning-tjänsten. Eftersom den bästa GPU-instansen som Azure erbjöd vid tidpunkten för den här studien omfattade åtta A100-GPU:er, beräknade vi kostnaderna för att köra fyra gånger så många GPU-instanser per uppgift för att ge en närmare uppskattning av prestandan de åtta H100-GPU:er som AWS-instanser och Dell PowerEdge XE9-servrar tillhandahåller. Vi använde sedan den timkostnaden för att avgöra hur mycket en användare skulle spendera för att köra varje instans för det förberäknade antalet timmar vi tog fram baserat på Dell-systemen. Efter att ha tillämpat den reserverade prissättningen för Azure under 3 år uppgick våra totala kostnader under 3 år till 2 231 805 USD. Fullständig information om instansen finns i tabell 15.

Tabell 15: Instanskostnader för Azure Machine Learning under en 3-årsperiod.

Tjänst	Instanstyp	Instans, USD/tim	Körtid (timmar/månad)	Kostnad för 3 år
Azure ML bärbara datorer	D2 v2	0,0476 USD	3 520	6 027,72 USD
Azure ML-bearbetning	M64	1,8258 USD	1 056	69 407,81 USD
Azure ML-träning	ND96amsr A100 v4	14 1475 USD	3 168	1 613 491,01 USD
Azure ML realtidsinferens	ND96amsr A100 v4	14 1475 USD	1 056	537 830,34 USD
Dataöverföringsåtgärder för Azure Block Blob Storage (10 000 000)				5 047,20 USD
Totalt (avrundat uppåt)				2 231 805 USD

1. AWS, "Get started with P5 instances", 29 april, 2024, https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/p5-instances-started.html.
2. Microsoft Build, "NDM A100 v4-series", 29 april 2024, https://learn.microsoft.com/en-us/azure/virtual-machines/ndm-a100-v4-series
3. Amazon, "Total Cost of Ownership of Amazon SageMaker", 29 april 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?", 20 april, 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu
och Hugging Face, "Model Memory Requirements", 29 april, 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", 29 april, 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", 29 april, 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.
7. The Jerusalem Post, "Maximizing Efficiency: Your 2023 Guide to GPU Servers", 8 april, 2024, https://www.jpost.com/insights/article-770858.
8. Systems Administrator II total compensation (salary and benefits) of $130,616 per year. Källa: Salary.com, "Systems Administrator II", 25 mars 2024, https://www.salary.com/tools/salary-calculator/systems-administrator-ii-benefits.
9. Dell, "The market's most complete deployment offer", 28 mars, 2024, https://www.delltechnologies.com/asset/en-us/services/deployment/briefs-summaries/prodeploy_plus_deployment_unification_ds.pdf.
10. Principled Technologies, "Using Dell ProDeploy Plus for Infrastructure can improve deployment times for Dell technology" 28 mars, 2024, https://www.delltechnologies.com/asset/en-us/products/cross-company/industry-market/principled-technologies-prodeploy-plus-for-infrastructure-services-whitepaper.pdf.
11. Dell, "Dell Enterprise Infrastructure Planning Tool", 24 mars 2024, https://dell-ui-eipt.azurewebsites.net/#/
12. Uptime Institute, "Large data centers are mostly more efficient, analysis confirms", 29 mars, 2024, https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/.
13. EIA, "Electricity Data Browser", 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", 25 mars, 2024, https://dell-ui-eipt.azurewebsites.net/#/.
15. VMware by Broadcom partner, Softchoice, uses this rack cost in a TCO analysis comparing costs of running VMware on-premises or in the cloud. Källa: Softchoose, "VMware Cloud on AWS", 24 mars 2024, https://www.softchoice.com/technology-partners/vmware/cloud-on-aws-tco-calculator.

Se den ursprungliga, engelska versionen av den här rapporten på https://facts.pt/9PHKEUe

Det här projektet har beställts av Dell Technologies.

Principled Technologies®

Facts matter.®