A Principled Technologies report: In-depth research. Real-world value.

Investing in GenAI: Cost-benefit analysis of Dell on-premises deployments vs. similar AWS and Azure deployments

In the tech world, Generative AI (GenAI) is the new frontier. As companies around the globe begin exploring how GenAI can further their business goals, they face numerous hurdles in implementing a GenAI solution that can address their specific needs. Among the biggest challenges are determining precisely how much a right-sized GenAI solution will cost and whether to deploy it on-premises or in the cloud. As one quote from Gartner analyst Frances Karamouzis put it: "Cost is one of the greatest threats to the success of AI and generative AI. More than half of the organizations are abandoning their efforts due to missteps in estimating and calculating costs."[1]

To provide organizations with a jumping-off point for understanding the total cost of deploying and managing GenAI workloads, including model fine-tuning and inferencing, we looked at the approximate 3-year costs of two on-premises Dell™ solutions leveraging PowerEdge™ R660 and PowerEdge XE9680 hardware—a traditional solution and a subscription-based Dell APEX pay-per-use solution—and comparable Amazon Web Services (AWS) SageMaker and Microsoft Azure Machine Learning solutions. According to our calculations, the Dell APEX pay-per-use solution was the most cost-effective of the 3-year solutions we compared. The competitive cloud solutions from AWS and Azure cost up to 3.81 times as much as the subscription-based Dell APEX pay-per-use solution. Compared to the traditional Dell on-premises solution, the AWS and Azure cloud solutions we priced would cost up to 2.88 times as much. Read on to see how GenAI can help your company and how we calculated our total cost of ownership (TCO) results.

Pay less for GenAI with a Dell APEX pay-per-use solution and a Dell on-premises GenAI solution

Competitive AWS and Azure cloud solutions cost more:

Up to 3.81x the cost of a Dell APEX pay-per-use solution

Up to 2.88x the cost of the traditional on-premises Dell solution

The benefits of using pre-trained models for Generative AI

Artificial intelligence (AI) models are systems that aim to mimic aspects of human intelligence or behavior. Companies as well as private individuals are using GenAI tools (such as ChatGPT and DALL-E) to generate content, including text, audio, videos, images, code, and simulations, as well as more complex outputs such as personalized marketing content, custom applications, and software.[2] A business looking to integrate generative AI into its operations can choose between using a pre-trained model or creating its own model from scratch. A pre-trained model, such as Llama 2, is already trained on a basic dataset for the model's intended use. From there, companies can fine-tune the model using their specific data, helping jumpstart the process of creating a model customized to their data and use cases.[3] According to one source, using pretrained models could cut the time to having a functional AI model by up to a year while saving hundreds of thousands of dollars.[4]

TCO scenario and solutions overview

New workloads often mean new investments. Many AI workloads require high-performance components in addition to the large amounts of storage already containing your data. Figuring out how to implement AI workloads involves balancing security, time, performance and scalability, ease of use, and cost. To provide an idea of how much AI solutions cost, we created an AI scenario using the open-source Llama 2 13B model and compared the cost to run the workload in four different environments. Our scenario included four specific tasks in a GenAI workload: data scientist coding and other work, data processing tasks, model fine-tuning tasks, and inferencing tasks. These tasks combine to keep the model accurate and up-to-date with the latest company-generated data to provide optimal model outputs. Table 1 shows the high-level specifications of the four environments we researched. Note: We completed all research and pricing on March 29, 2024, with prices subject to change after this date.

Table 1: Solution details for the TCO comparison.

Task	Server/instance	GPUs per server/instance	Additional purchases
Traditional on-premises solution			
Cluster management	3x PowerEdge R660	N/A	2x PowerSwitch S5232-ON Network Infrastructure and 1x PowerSwitch N3200-ON OOB Management
Notebooks			
Data processing	2x PowerEdge XE9680	8x NVIDIA H100	
Model fine-tuning			
Inference			
Managed on-premises Dell APEX pay-per-use solution			
Cluster management	3x PowerEdge R660	N/A	2x PowerSwitch S5232-ON Network Infrastructure and 1x PowerSwitch N3200-ON OOB Management
Notebooks			
Data processing	2x PowerEdge XE9680	8x NVIDIA H100	
Model fine-tuning			
Inference			

Task	Server/instance	GPUs per server/instance	Additional purchases
AWS SageMaker solution			
Cluster management	N/A	N/A	7TB EBS storage per month for ml.r5.16xlarge instances and 1TB in and 15TB out S3 data transfer
Notebooks	20x ml.t3.medium	N/A	
Data processing	2x ml.r5.16xlarge	N/A	
Model fine-tuning	ml.p5.48xlarge	8x NVIDIA H100	
Inference	ml.p5.48xlarge	8x NVIDIA H100	
Azure Machine Learning solution			
Cluster management	N/A	N/A	10,000,000 Azure Block Blob Storage data transfer operations
Notebooks	20x D2 v2	N/A	
Data processing	M64	N/A	
Model fine-tuning	4x ND96amsr A100 v4	8x NVIDIA A100	
Inference	4x ND96amsr A100 v4	8x NVIDIA A100	

For exact specifications of the solutions we compared, see the science behind the report.

For this analysis, we tried to create a broadly applicable example scenario to estimate cost differences across environments. We chose the Llama 2 13B GenAI model because it is a widely available, open-source model. We included costs for data scientists' machine learning development notebooks, data processing tasks, continuous model fine-tuning, and real-time inference. We did not include costs for storage beyond that which the servers or instances needed to do their tasks.

For the on-premises Dell solutions, we assumed the dev notebooks and cluster management tasks would take place on the Dell PowerEdge R660 cluster, while the processing, fine-tuning, and inference tasks would take place on the Dell PowerEdge XE9680 cluster.

For the cloud solutions, we chose instances to fit a task's needs; notebook instances were very small, while we gave processing instances significant memory. Because the public cloud services spin up a new instance for each task, each of these tasks would have a dedicated eight-GPU instance for its run duration. Thus, we calculated the number of tasks the PowerEdge XE9680 servers could perform while maintaining the same GPU-per-task ratio. We also added an estimate for the costs of data transfer to and from the cloud provider's object storage to account for the cost of moving data through the cloud.

To account for varying business realities and make a fair comparison, we made the following assumptions:

- All costs exclude taxes, as specific rates vary by location.
- All software is open source, with licenses allowing commercial usage.
- We exclude management costs for the cloud solutions. For the on-premises solutions, we factor in ongoing system administration costs to maintain the hardware and support the data scientists.
- For the on-premises solutions, we consider costs for physical data center space and power and cooling; we factor these costs into instance costs for the cloud solutions.

For more details of our assumptions and calculations, see the science behind the report.

Comparing the costs for GenAI: On-premises Dell solutions vs. the cloud

Assumptions for GenAI cost comparisons

- We assume there are 22 workdays in each month, with workloads set to run overnight to maximize usage.
- Thus, each server offers 528 hours of runtime per month.
- Data processing tasks can run the full 528 hours x two Dell PowerEdge XE9680 servers = 1,056 hours runtime.
- Twenty data scientists work 8 hours a day for 22 days a month for a total of 3,520 hours.

Since the processing tasks use the CPU and memory, we host them for the full 1,056 server uptime hours on the PowerEdge XE9680 servers. We split the model fine-tuning and inferencing tasks between the two servers with the assumption that the workload would require more fine-tuning time than inferencing time. Thus, we calculated 792 hours per month spent on fine-tuning tasks and 264 hours per month on inferencing tasks.

Finally, for the 20 developers' notebook usage, we assumed each had a typical 8-hour workday for 5 days a week, totaling 3,520 hours per month. The number of data scientists your company employs to maintain and fine-tune your model will depend on several factors such as how many different ways you want to interpret your data set or how many applications your data set feeds. We chose a number on the higher end of the scale to represent an up-to cost that would apply to many companies. Since these instances in the public cloud are very small and cost very little relative to the solution as a whole, the number of data scientists will not have a large impact on the total cost of our solution. Using these uptime calculations, we were able to plug in the number of hours each instance type would run per month on the two cloud solutions. For the final total costs of all solutions, see the science behind the report.

Pricing details for the traditional on-premises Dell solution

We contacted Dell and asked for a Dell Recommended Price quote for our traditional on-premises solution. This quote included the cost of servers and switches, ProDeploy Plus for on-site installation services for the servers, and a 5-year ProSupport for Infrastructure plan to provide support and maintenance services for the gear. Note: We opted for a 5-year support plan because while we limited our TCO to 3 years, most servers last 3 to 5 years and need service beyond the three years we looked at. We then calculated the power and cooling energy costs and data center rack space costs for a period of 3 years, as well as the administrative costs for maintaining the gear for 3 years.

Pricing details for the AWS SageMaker cloud solution

AWS breaks down its SageMaker service into several subservices covering tasks such as processing and training as well as data scientists' notebooks. Note that while we are fine-tuning a pre-trained model, the AWS SageMaker subservice is called SageMaker Training. To obtain SageMaker pricing, we used the AWS Pricing Calculator and the Machine Learning Savings Plans calculator.[5,6] For our TCO, we priced instances for notebooks, processing, model fine-tuning, and inference as follows:

Table 2: AWS SageMaker environment instances and run time hours per month.

Instance model	# of instances	Task	Run time (hours/month)/ instance
ml.t3.medium	20	Data scientist notebook	176
ml.r5.16xlarge	2	Data processing	1,056
ml.p5.48xlarge	1	Model fine-tuning	792
ml.p5.48xlarge	1	Inferencing	264

Pricing details assumptions:

- We chose two ml.r5.16xlarge instances for data processing to ensure at least 1 TB of memory per task based on research that indicated processing tasks are memory intensive.[7,8]
- We added 7 TB per month of EBS storage to the ml.r5.16xlarge instances as they do not come with disks.
- While we didn't estimate the costs of the storage hosting the main dataset, we did estimate S3 data transfer costs for 1 TB in and 15 TB out per month to account for the subsets of data the training and inference tasks will be using.
- The ml.p5.48.large instances came equipped with direct-attached NVMe storage, so we did not add EBS storage for those instances.

Note: SageMaker includes an Elastic Fabric Adapter (EFA) that offers high throughput rates.[9] While we believe the networking in the Dell solution is adequate for our scenario, you could opt to purchase a network configuration with more bandwidth. Thus, it's possible that the AWS solution could process more tasks than the Dell solution depending on your networking choices.

AWS offers both on-demand pricing and SageMaker savings plans. On-demand pricing is the most expensive, while the savings plans offer up to 64 percent reduced costs with a 3-year commitment.[10] We priced the AWS configuration using the 3-year commitment price.[11] In addition, AWS offers customers the option to pay costs upfront for a greater cost reduction, which we chose to do for our TCO calculations.

Comparing the traditional, on-premises Dell solution to AWS SageMaker

Using the above assumptions for both solutions, we calculated a 3-year TCO comparison. Our calculations show that choosing the traditional, on-premises Dell PowerEdge solution to run GenAI workloads could offer real savings compared to running the same workload on AWS SageMaker.

As Figure 1 shows, we calculated that the AWS SageMaker solution could cost up to 2.88 times as much as the on-premises Dell solution. Note: For this and all subsequent cost details, see the science behind the report.

AWS SageMaker could cost nearly 3x as much as a similar Dell solution

Nearly break even at 1 year

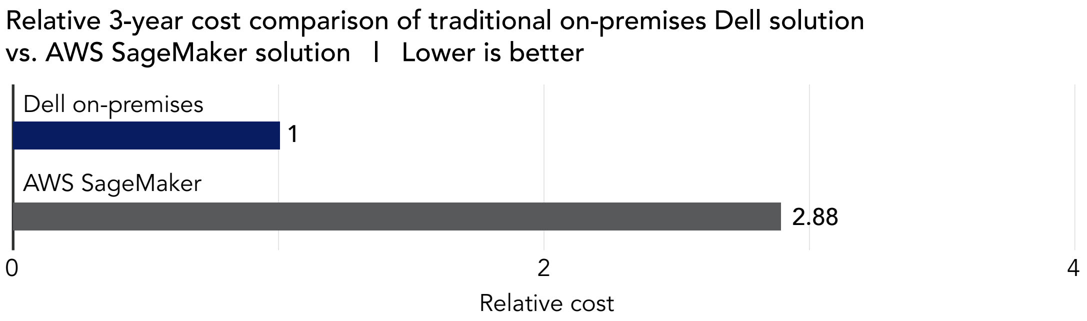

Figure 1: Relative costs of a GenAI Dell on-premises solution and an AWS SageMaker solution over 3 years.

Given the over 2.8x higher cost over 3 years, users can assume that even with the costs for hosting, cooling, and managing an on-premises solution, they would nearly break even at 1 year compared to the cost of AWS hosting.

Pricing details for the Azure Machine Learning cloud solution

For the Azure Machine Learning service environment, we chose instances for the same four tasks as the AWS environment: data scientist developer notebooks, data processing, fine-tuning, and inference. We obtained our pricing from the Azure Pricing Calculator, choosing the 3-year reserved savings plan option.[12] The instances we priced are as follows:

Table 3: Azure Machine Learning environment instances and run time hours per month.

Instance model	# of instances	Task	Run time (hours/ month/ instance)
D2 v2	20	Data scientist notebook	176
M64	1	Data processing	1,056
ND96asmr A100 v4	4	Model fine-tuning	792
ND96asmr A100 v4	4	Inferencing	264

Price details for Azure Machine Learning assumptions

- The Azure Machine Learning service did not offer an instance with the NVIDIA H100 GPUs, so we chose four A100 GPU instances to approximate similar performance.[13]
- All Azure Machine Learning instances come with attached block storage, so we did not price additional storage for the Azure environment. As in our AWS calculations, however, we did approximate 10,000,000 Block Blob Storage data transfer operations for transferring data into and out of the Machine Learning instances.

Azure offers pay-as-you-go pricing, Azure savings plans, and Azure Reservations options for the Machine Learning service.[14] Azure did not offer a cheaper pay up front option, as AWS did. To best match how we priced the AWS environment, we opted for the 3-year Reservations plan pricing.

Comparing the on-premises Dell solution to Azure ML

Using the above assumptions, we calculated the costs of a 3-year Azure solution and compared it to our 3-year TCO estimates for the on-premises Dell solution. Again, our calculations show that the traditional, on-premises Dell PowerEdge solution for GenAI workloads can offer significant 3-year savings over a comparable Azure ML solution.

In fact, we estimate that the total costs of the Azure Machine Learning solution over 3 years would be 2.72 times as much as the traditional on-premises Dell solution (see Figure 2). These results show that keeping your hardware in house for GenAI with a traditional Dell solution can help make your GenAI budget reasonable so you can use those savings to innovate elsewhere.

Azure Machine Learning could cost nearly 3x as much as a similar Dell solution

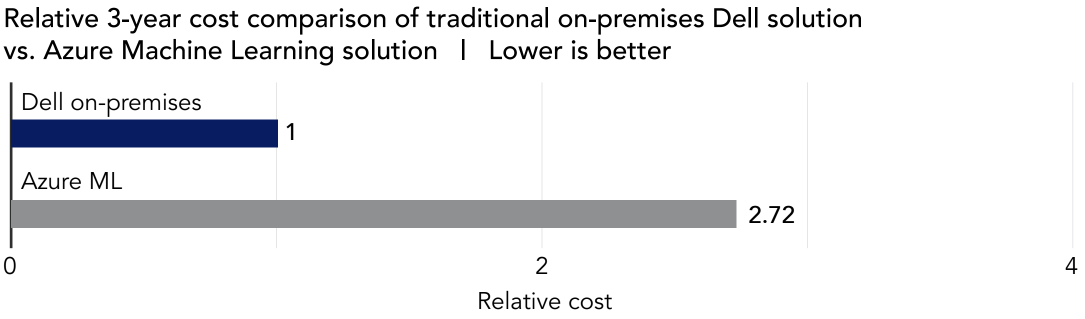
Relative 3-year cost comparison of traditional on-premises Dell solution vs. Azure Machine Learning solution | Lower is better

Figure 2: Relative costs of a GenAI Dell on-premises solution and an Azure Machine Learning solution over 3 years.

Similar to the AWS solution, at over 2.7x the cost over 3 years, Dell on-premises customers could expect to come close to breaking even at 1 year compared to the Azure pricing.

Save even more by subscribing to a Dell APEX pay-per-use solution

Some organizations may find the long-term commitment inherent in a traditional on-premises solution prohibitive. That's why Dell offers a Dell APEX pay-per-use solution. Dell can install hardware in your organization's data center, so it remains on premises like the traditional solution, and offers a 3-, 4-, or 5-year commitment using a Dell APEX pay-per-use solution for compute resources at a specified consumption rate for a consistent monthly payment. If you need more than your committed consumption level, you can tap into the remaining resources for an additional cost. When your subscription ends, you can cancel the service and return the hardware, renew as-is, or switch to a solution that better fits your needs at the time.[15]

For our TCO comparison, we received a quote from Dell for the same hardware we included in our traditional on-premises environment, but also adding a 3-year subscription to a Dell APEX pay-per-use solution at a 75 percent guaranteed consumption rate. The Dell APEX pay-per-use solution consumption rates for servers are based on the amount of time a server spends at greater than 5 percent CPU activity in a month. Our assumptions were as follows:

Dell APEX pay-per-use solution assumptions

- Roughly 726 hours per month with a 75 percent guaranteed consumption rate = maximum of 544.5 hours of server time per month before needing additional resources. For consistency with the other calculations, we used 528 hours per month.
- The quote also included ProDeploy Plus and ProSupport Next-Business Day plans, so we did not include admin costs for initial setup.
- We included the same power and cooling and data center rack space costs as our traditional solution.

We found that a Dell APEX pay-per-use solution, which combines the security and control advantages of a traditional on-premises solution with the convenience and flexibility of a managed service, could save organizations a significant amount over 3 years, compared to the cloud solutions that we priced.

As Figure 3 shows, the AWS SageMaker solution could cost 3.81 times as much as the Dell APEX pay-per-use solution.

AWS SageMaker could cost nearly 4x as much as a similar Dell APEX pay-per-use solution

Break even in under 1 year

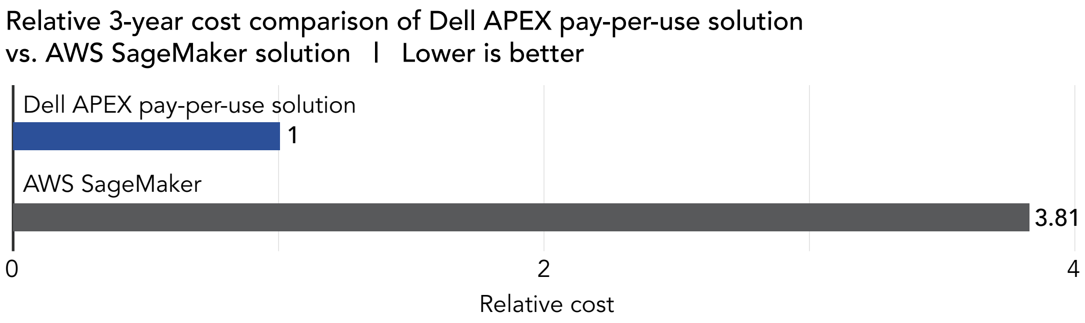

Figure 3: Relative costs of a GenAI Dell APEX pay-per-use solution and an AWS SageMaker solution over 3 years.

Because it costs over 3.8 times more than a Dell APEX pay-per-Use Solution over 3 years, users can assume that they would break even before even 1 year is up. The cost savings of a Dell APEX pay-per-use solution were similar for the Azure Machine Learning solution, which cost 3.60 times as much as the Dell APEX pay-per-use solution (see Figure 4).

Azure Machine Learning could cost nearly 4x as much as a similar Dell APEX pay-per-use solution

Break even in under 1 year

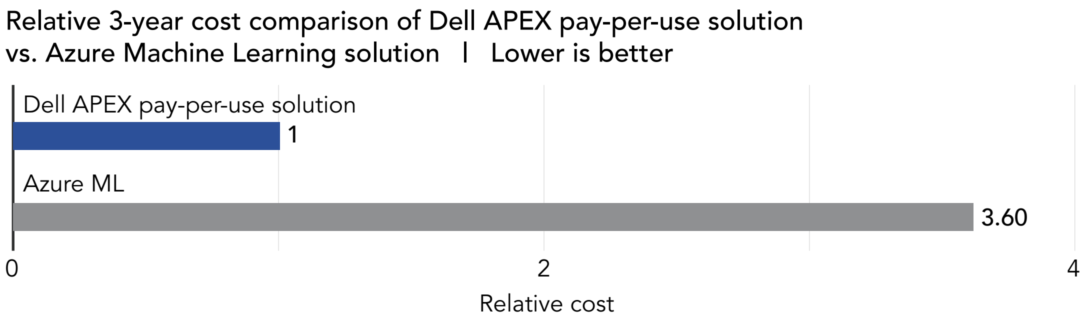

Figure 4: Relative costs of a GenAI Dell APEX pay-per-use solution and an Azure Machine Learning solution over 3 years.

These results show that budget-conscious organizations seeking to implement GenAI could meet their needs well by selecting an on-premises Dell APEX pay-per-use solution rather than hosting these potentially sensitive workloads in the cloud. In addition, as with the AWS comparison, customers can likely assume that they would break even before 1 year is up compared to the Azure solution.

Additional considerations for GenAI workloads

Other benefits of choosing on-premises vs. Cloud

While cloud hosting offers benefits such as scalability and flexibility, there are several concerns besides cost you should consider before hosting your LLMs on a public cloud. One of the biggest is the risk associated with hosting large amounts of user data on a third-party platform. Many LLMs collect user data to improve their models, and that data must reside somewhere for the models to access it. Storing this sensitive data in the cloud can expose it to risks such as:

- Exposing data to public interfaces that attackers might access. For example, CrowdStrike discovered one such vulnerability that allowed them to find AWS S3 buckets based on DNS requests.[16]
- Multiplying complexity that could lead to misconfigurations as your IT teams juggle multiple services and cloud providers that change defaults and settings regularly.
- Magnifying human error when using cloud-based APIs that could expose sensitive data.[17]

Private LLMs reduce these risks, as users typically have greater control over their data centers and thus their data streams, network isolation, API controls, and more. Furthermore, users running LLMs locally have more control over the entire stack, from the hardware the LLM runs on to the model and data enabling the solution. Admins can use additional training to ensure that local LLMs comply with specific regulations. In the cloud, users have less control over the underlying infrastructure and implementation.[18] Additionally, on-premises solutions can keep costs predictable instead of varying month to month.

Data storage and transfer are a big part of the LLM application requirements. Training an LLM requires large amounts of data, which must be stored and then transferred from storage to compute resources for processing. If the devices, databases, and user data feeding your LLM are already storing their data on premises, the costs of transitioning that data to the cloud and the network bandwidth needed could be high.

The Dell AI portfolio

Clearly, implementing AI workloads requires more than just determining the overall costs of the solution. You must determine which model is best for your needs and design a solution that can handle the amount of data you have. After selecting the right model and deciding on a solution comes staffing concerns. You need to train or hire staff to ensure your IT staff understand data science and can properly train and execute your AI solution.

Dell offers a complete AI portfolio that does more than give you cost savings on hardware: they provide the tools you need to overcome challenges during your AI implementation. The Dell AI portfolio spans hardware, data management services, professional services, AI training, reference architectures, and many partnerships with other AI-focused companies.[19] Dell can help you design, plan for, and implement a successful AI solution fit for your needs. To read more about how the Dell AI portfolio stacks up against competitor offerings, you can read our reports comparing the Dell AI portfolio to those from Supermicro[20] and HPE.[21]

Llama 2 models

Llama 2, which stands for Large Language Model Meta AI, is a free and versatile language processing technology developed by Meta. It is a pretrained large language model (LLM) with three main model size variants based on the number of parameters (7B, 13B, and 70B), each providing different performance characteristics.[22] Meta trained Llama 2 with a reinforcement learning approach to produce family-friendly output to users, with the aim to become familiar with human choices and preferences.[23] Read more about Llama 2 at https://llama.meta.com/llama2/.

Conclusion

Diving into the world of GenAI has the potential to yield a great many benefits for your organization, but it first requires consideration for how best to implement those GenAI workloads. Whether your AI goals are to create a chatbot for online visitors, generate marketing materials, aid troubleshooting, or something else, implementing an AI solution requires careful planning and decision-making. A major decision is whether to host GenAI in the cloud or keep your data on premises. Traditional on-premises solutions can provide superior security and control, a substantial concern when dealing with large amounts of potentially sensitive data. But will supporting a GenAI solution on site be a drain on an organization's IT budget?

In our research, we found that the value proposition is just the opposite: Hosting GenAI workloads on premises, either in a traditional Dell solution or using a managed Dell APEX pay-per-use solution, could significantly lower your GenAI costs over 3 years compared to hosting these workloads in the cloud. In fact, we found that a comparable AWS SageMaker solution would cost up to 3.8 times as much and an Azure ML solution would cost up to 3.6 times as much as GenAI on a Dell APEX pay-per-use solution. These results show that organizations looking to implement GenAI and reap the business benefits to come can find many advantages in an on-premises Dell solution, whether they opt to purchase and manage it themselves or choose a subscription-based Dell APEX pay-per-use solution. Choosing an on-premises Dell solution could save your organization significantly over hosting GenAI in the cloud, while giving you control over the security and privacy of your data as well as any updates and changes to the environment, and while ensuring your environment is managed consistently.

[1] LinkedIn, Jacquecine Burrell's post, accessed April 19, 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," accessed April 3, 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," accessed April 3, 2024, https://oneai.com/learn/pre-trained-ai-model.

[4] NVIDIA, "What Is a Pretrained AI Model?" accessed April 3, 2024, https://blogs.nvidia.com/blog/what-is-a-pretrained-ai-model/.

[5] AWS, "AWS Pricing Calculator," accessed April 3, 2024, https://calculator.aws/#/.

[6] AWS, "Machine Learning Savings Plans," accessed April 3, 2024, https://aws.amazon.com/savingsplans/ml-pricing/.

[7] StackOverflow, "Why should preprocessing be done on CPU rather than GPU?" accessed April 3, 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu.

8. Hugging Face, "Model Memory Requirements," accessed April 3, 2024, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.

9. AWS, "Training large language models on Amazon SageMaker: Best practices," accessed April 3, 2024, https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/.

10. AWS, "Machine Learning Savings Plans," accessed April 3, 2024, https://aws.amazon.com/savingsplans/ml-pricing/.

11. Note: AWS confirmed that the ml.p5.48xlarge instance is included in the 3-year commitment price plan. At the time of this study, it was not listed in the savings plan calculator. We estimated the cost of the ml.p5 instance by using the savings percentage listed for the non-machine-learning version p5.48xlarge as listed at https://aws.amazon.com/savingsplans/compute-pricing/.

12. Microsoft, "Azure Pricing Calculator," accessed April 3, 2024, https://azure.microsoft.com/en-us/pricing/calculator/.

13. Comet, "Comparison of NVIDIA A100, H100 + H200 GPUs," accessed April 3, 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.

14. Microsoft, "Azure Machine Learning pricing," accessed April 3, 2024, https://azure.microsoft.com/en-us/pricing/details/machine-learning/.

15. Dell, "Dell APEX Flex on Demand," accessed April 3, 2024, https://www.delltechnologies.com/partner/en-us/partner/apex-flex-on-demand.htm.

16. CrowdStrike, "12 Cloud Security Issues: Risks, Threats, and Challenges," accessed April 3, 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," accessed April 3, 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," accessed April 29, 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," accessed April 29, 2024, https://facts.pt/q9p46K9.

21. Principled Technologies, "Meeting the challenges of AI workloads with the Dell AI portfolio," accessed April 29, 2024, https://facts.pt/zPmSx4c.

22. Meta, "Llama 2: open source, free for research and commercial use," accessed April 3, 2024, https://llama.meta.com/llama2/.

23. Pavan Belagatti, "Unpacking Meta's Llama 2: The Next Leap in Generative AI," accessed April 3, 2024, https://www.singlestore.com/blog/a-complete-beginners-guide-to-llama2/

Read the science behind this report at https://facts.pt/EtDSF2Z ▶

Principled Technologies®

Facts matter.®

Principled Technologies is a registered trademark of Principled Technologies, Inc.
All other product names are the trademarks of their respective owners.
For additional information, review the science behind this report.

This project was commissioned by Dell Technologies.