Make GenAI investments go further with the Dell AI Factory

The cost benefit of implementing a Dell AI Factory solution versus AWS and Azure

Generative AI (GenAI) can help organizations of all types and sizes further their business goals, but selecting a right-sized solution that meets performance, budget, and security needs can be a challenge. While it may appear wise to host large language models (LLMs) in the cloud to maximize flexibility, committing to a deploying AI outside your data center walls can lead to budget concerns over time and costing businesses more in the long run. In a 2024 survey, 46 percent of business leaders said AI implementation cost was a concern—a significant jump from just 3 percent the previous year. Companies are beginning to halt or postpone AI initiatives due to costs, as they incur "token costs, unexpected additional costs, and AI sprawl."[1]

To help businesses understand the total cost of deploying and managing GenAI workloads, including model fine-tuning and inferencing, we looked at the approximate 4-year costs of an on-premises Dell™ AI Factory solution with PowerEdge™ R660 and PowerEdge XE9680 hardware using two payment options—with a traditional payment model CAPEX solution and a Dell APEX Subscriptions solution—and comparable Amazon Web Services (AWS) SageMaker and Microsoft Azure Machine Learning solutions. According to our calculations, the Dell AI Factory solutions were the most cost-effective of the 4-year solutions we compared. The Dell APEX Subscriptions solution reduced costs by 71 percent compared to AWS and 60 percent compared to Azure. The same Dell AI Factory on-premises solution with no subscription (CAPEX model) would reduce 4-year costs by 71 percent compared to the AWS solution and 61 percent compared to the Azure cloud solution we priced. Read on to see how choosing to run GenAI on premises with an on-premises Dell AI Factory solution can help your company make the most of your investment.

Get more for your investment with a solution from the Dell AI Factory

Save up to 71%
vs. a competitive AWS solution

Break even at 1 year

Save up to 61%
vs. a competitive Azure solution

Break even at 1.5 years

4-year TCO costs for a Dell AI Factory on-premises solution vs. AWS and Azure environments | *Lower is better*

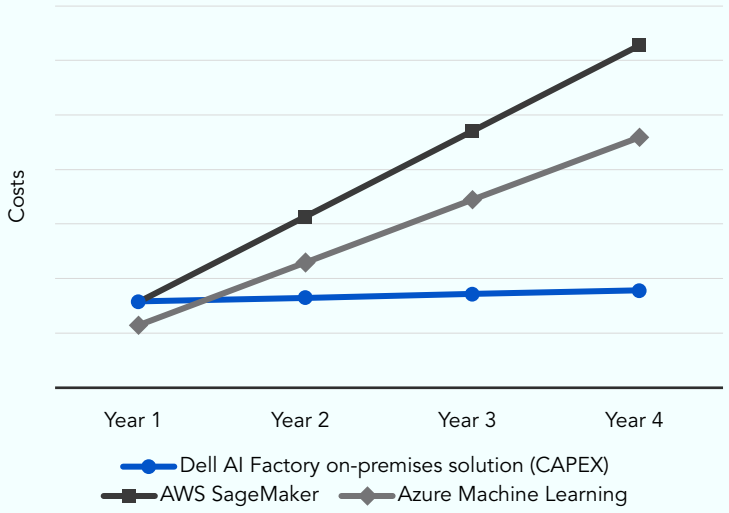

Figure 1: Relative costs of a Dell AI Factory on-premises solution (CAPEX) vs. AWS SageMaker and Azure Machine Learning solutions over 4 years.

About the Dell AI Factory

The Dell AI Factory is a comprehensive approach built to meet the shifting demands of modern AI applications. Powered by Dell, the AI Factory combines cutting-edge infrastructure, software, and services, offering flexibility to accelerate AI initiatives. Organizations can adopt the entire solution, including all components for a streamlined approach, or customize their setup by leveraging Dell infrastructure and services tailored for specific GenAI outcomes. The open-ecosystem approach ensures that Dell AI Factory solutions are compatible with diverse technologies and workflows.

The Dell AI Factory provides a wide range of infrastructure options to meet modern enterprise needs, including the option to engage with Dell APEX Subscriptions, making it a valuable framework for companies seeking to leverage AI to transform their data into positive business outcomes. To learn more about what the Dell AI Factory has to offer your organization, visit dell.com/ai.

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 3 8B 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 machine learning development 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 27, 2025, 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
Dell AI Factory 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 (each with 30TB storage)	8x NVIDIA H100	
Model fine-tuning			
Inference			
Managed on-premises Dell APEX Subscriptions 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 (each with 30TB storage)	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	1x ND96isr H100 v5	8x NVIDIA H100	
Inference	1x ND96isr H100 v5	8x NVIDIA H100	

Please note that this study uses pricing for NVIDIA H100 GPUs. While NVIDIA has released H200 GPUs, and PowerEdge XE9680 servers support them, we opted to compare TCO for similarly configured solutions using H100 GPUs. 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 3 8B 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 development 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 geographic 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.
- For the Dell AI Factory CAPEX purchase, we excluded any working cost of capital/depreciation calculations.

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

Comparing the costs for GenAI: Dell AI Factory on-premises solutions vs. the cloud

Assumptions for GenAI cost comparisons

- We assume there are 22 workdays in each month, with workloads set to run for 24 hours 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 analysts work 8 hours a day for 22 days a month for a total of 3,520 hours.

Since the processing tasks use 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 data scientists' 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 Dell AI Factory on-premises solution

Dell provided a quote using the Dell Recommended Price quote for the Dell AI Factory 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 4 years, most servers last 3 to 5 years and need service beyond the 4 years we looked at. We then calculated the power and cooling energy costs and data center rack space costs for a period of 4 years, as well as the administrative costs for maintaining the gear for 4 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.[2,3] 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.[4,5]

- We added 3.5 TB per month of EBS storage to each 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.[6] 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.[7] AWS does not offer specific pricing for a 4-year commitment; therefore, to work based on the best possible cost for our 4-year TCO, we priced the AWS configuration using the 3-year commitment price prorated to 4 years.[8] (For an alternate look at AWS pricing, we also calculated 4-year costs using 3 years at the 3-year commitment price plus 1 year at the 1-year commitment price. See the science report for these additional results.) 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. Note that we priced our AWS solution in the US East (Ohio) region, and that pricing may vary by region.

Spend less with a Dell AI Factory on-premises solution compared to AWS SageMaker

Using the above assumptions for both solutions, we calculated a 4-year TCO comparison. Our calculations show that choosing the Dell AI Factory on-premises solution to run GenAI workloads could offer real savings compared to running the same workload on AWS SageMaker.

As Figure 2 shows, we calculated that the Dell AI Factory on-premises solution would cost 71 percent less than a similar AWS SageMaker solution. Given that the AWS solution is 3.5x the cost over 4 years, users can assume that they would nearly break even at 1 year with a Dell AI Factory on-premises solution compared to the cost of AWS hosting.

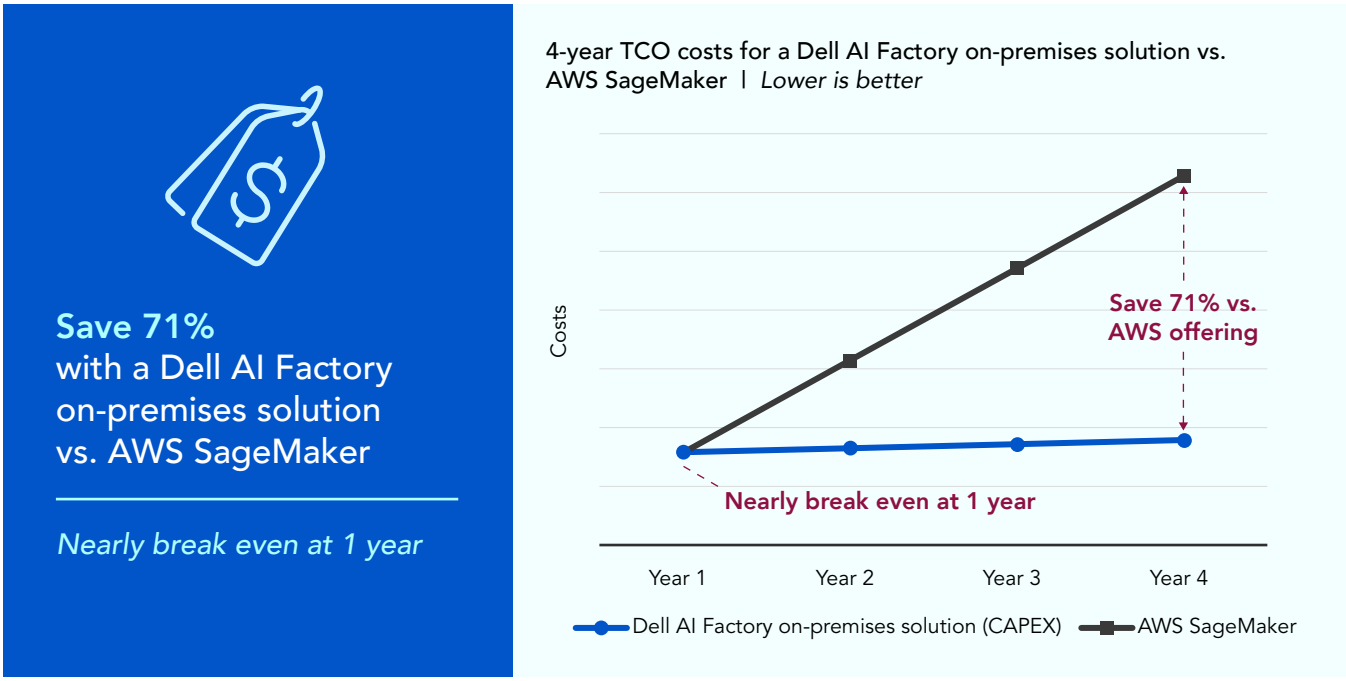

Figure 2: Relative costs of a Dell AI Factory on-premises solution and an AWS SageMaker solution over 4 years.

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 4-year reserved savings plan option.[9] 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
ND96isr H100 v5	1	Model fine-tuning	792
ND96isr H100 v5	1	Inferencing	264

Price details for Azure Machine Learning assumptions

- 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. The calculator for these transactions includes several specific transactions, such as Write operations, Read operations, and more. We chose 10,000,000 for each.

Azure offers pay-as-you-go pricing, Azure savings plans, and Azure Reservations options for the Machine Learning service.[10] While Azure does offer a pay up front option, as AWS did, it did not appear to change the monthly cost or provide a discount. To best match how we priced the AWS environment, we opted for the 3-year Reservations plan pricing and calculated 4 years prorated. (As with AWS, we also calculated the costs and savings for a 3-year reserved plus 1-year reserved pricing, which you can see in the science behind this report.) As we did with AWS, we priced our Microsoft Azure solution in the East US 2 region. Please note that pricing may vary by region.

Break even in less than two years with a Dell AI Factory on-premises solution instead of Azure ML

Using the above assumptions, we calculated the costs of a 4-year Azure solution and compared it to our 4-year TCO estimates for the Dell AI Factory on-premises solution. Again, our calculations show that the Dell AI Factory on-premises solution for GenAI workloads can offer significant 4-year savings over a comparable Azure ML solution.

In fact, we estimate the Dell AI Factory on-premises solution costs 61 percent less than a similar Azure Machine Learning solution (see Figure 3). These results show that keeping your hardware in house for GenAI with a Dell AI Factory on-premises solution can help make your GenAI budget more reasonable. Because the Azure ML is over 2.5x the cost over 4 years, Dell AI Factory on-premises solution customers could expect to come close to breaking even at about a year and a half compared to the Azure pricing.

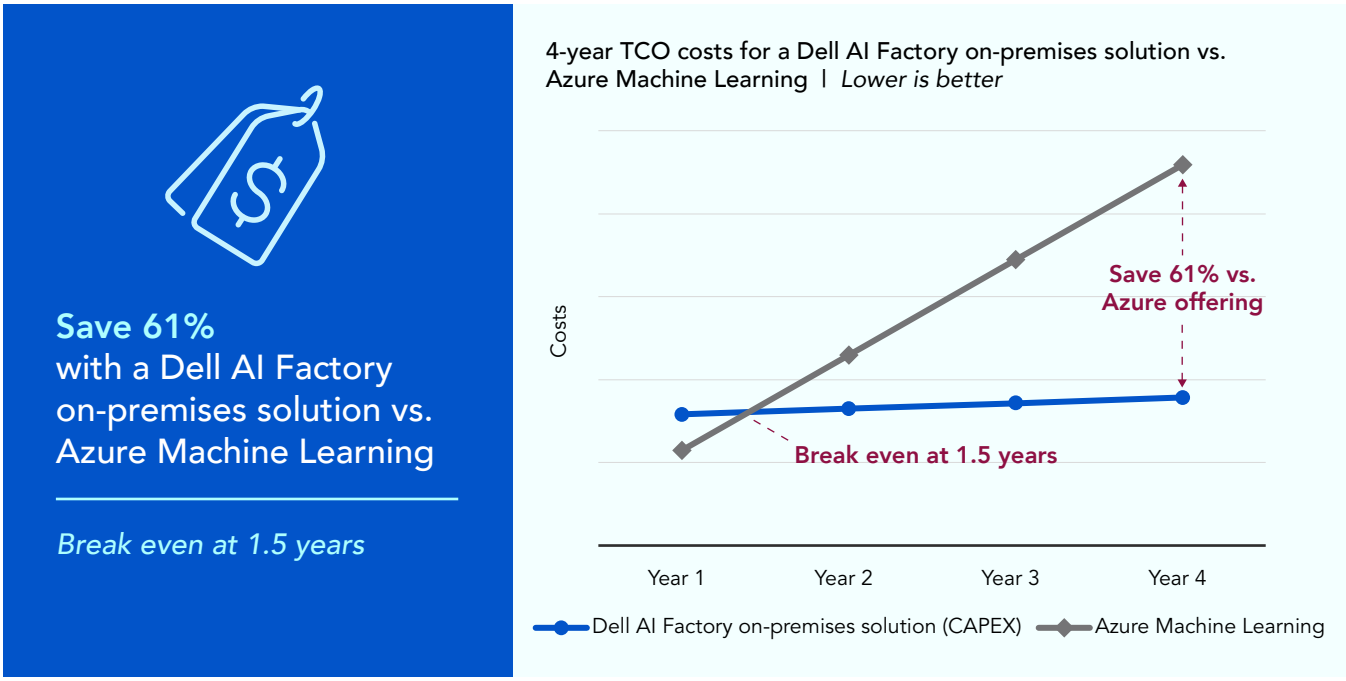

Figure 3: Relative costs of a Dell AI Factory on-premises solution and an Azure Machine Learning solution over 4 years.

Save by choosing Dell APEX Subscriptions

Some organizations may find the long-term commitment inherent in a traditional on-premises solution prohibitive. That's why Dell offers Dell APEX Subscriptions. 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 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.[11]

For our TCO comparison, we received a quote from Dell for the same hardware we included in our CAPEX Dell AI Factory on-premises solution, but also adding a 4-year subscription to Dell APEX Subscriptions at a 75 percent guaranteed consumption rate. The Dell APEX Subscriptions consumption rates for servers are based on the amount of time a server spends at greater than 5 percent CPU activity in a month.

Dell APEX Subscriptions 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 Dell APEX Subscriptions, 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 4 years, compared to the cloud solutions that we priced.

As Figure 4 shows, Dell APEX Subscriptions costs 71 percent less than the AWS SageMaker solution. The AWS solution costs 3.5 times more than Dell APEX Subscriptions over 4 years; with Dell APEX Subscriptions, you can pay less starting on day one and ultimately spend significantly less over 4 years.

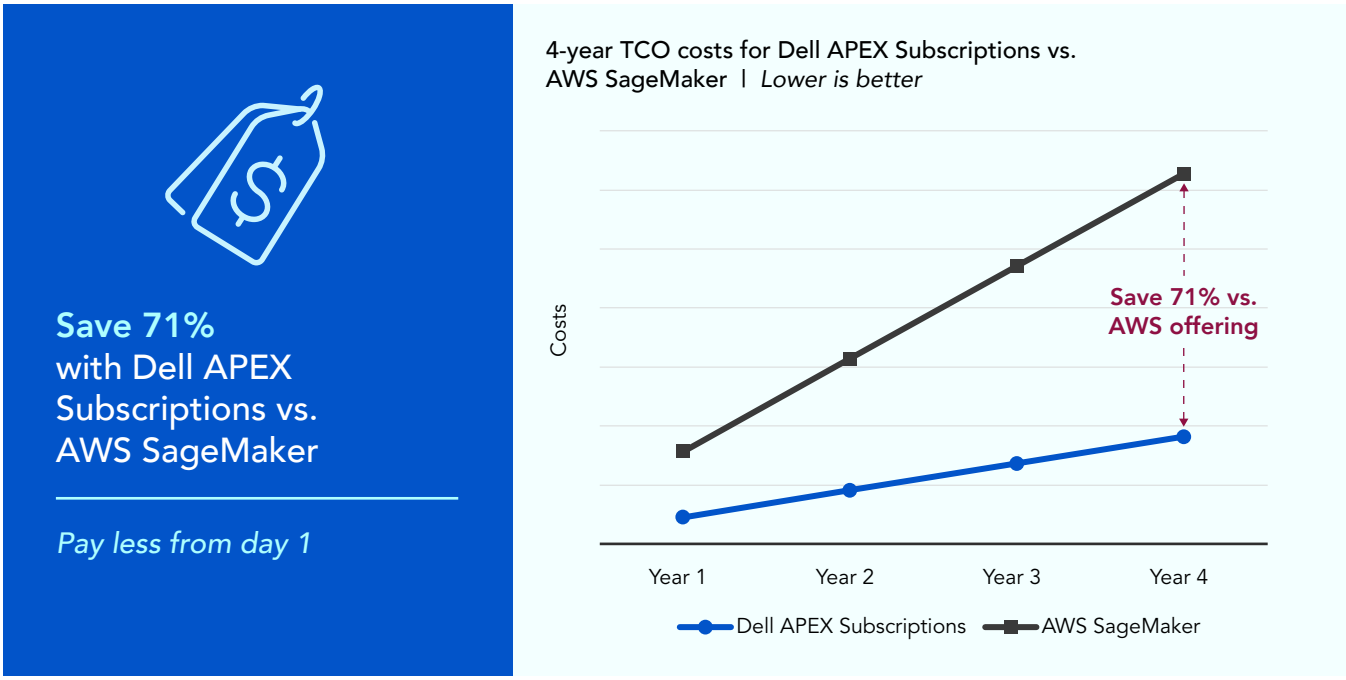

Figure 4: Relative costs of Dell APEX Subscriptions and an AWS SageMaker solution over 4 years.

As Figure 5 shows, Dell APEX Subscriptions offered significant cost savings compared to the Azure Machine Learning solution as well, reducing 4-year TCO by 60 percent. This means that Azure Machine Learning solution would cost 2.5 times as much as using Dell APEX Subscriptions over 4 years. These results show that budget-conscious organizations seeking to implement GenAI could meet their needs well by using Dell APEX Subscriptions rather than hosting these potentially sensitive workloads in the cloud. In addition, as with the previous comparison, customers pay less from the first day they engage Dell APEX Subscriptions, culminating in dramatically lower costs over the 4-year term.

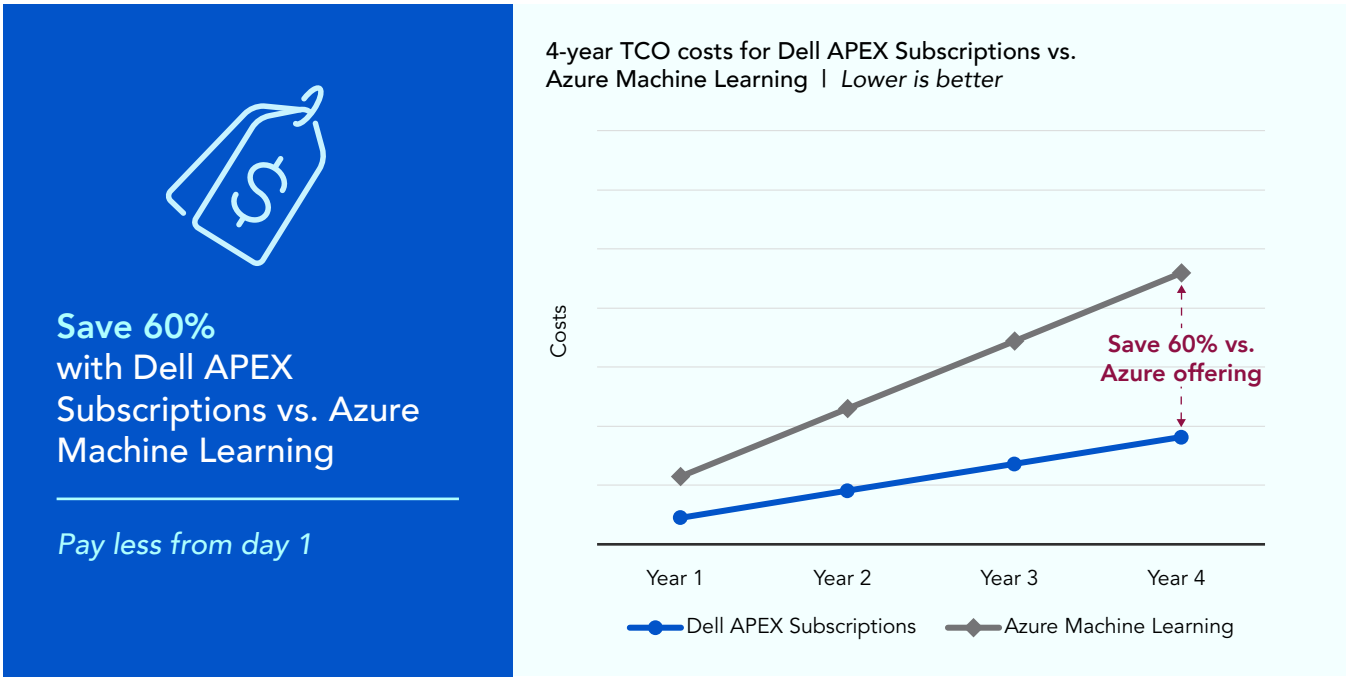

Figure 5: Relative costs of Dell APEX Subscriptions and an Azure Machine Learning solution over 4 years.

Additional considerations for running GenAI workloads on premises vs. the cloud

Placing large amounts of user data in the public cloud for LLMs to collect and refine on third-party platforms can present significant security risks, including:

- 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.[12]
- 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.[13]

Keeping LLMs on private networks can mitigate these risks, as in-house solutions have greater control over data streams, network isolation, API controls, maintaining data compliance, optimizing performance, 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.[14] 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 that must reside somewhere, and move between storage and 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.

Llama 3 models

Llama 3, 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 two main model size variants based on the number of parameters (8B and 70B), that support a wide range of use cases.[15] Meta trained Llama 3 with a "…new high-quality human evaluation set. This evaluation set contains 1,800 prompts that cover 12 key use cases: asking for advice, brainstorming, classification, closed question answering, coding, creative writing, extraction, inhabiting a character/persona, open question answering, reasoning, rewriting, and summarization."[16]

Read more about Llama 3 at https://ai.meta.com/blog/meta-llama-3/.

Conclusion

Our research shows that hosting GenAI workloads on premises, either in a traditional Dell solution or using managed Dell APEX Subscriptions, could significantly lower your GenAI costs over 4 years compared to hosting these workloads in the cloud. In fact, we found that a Dell AI Factory on-premises solution could reduce costs by at much as 71 percent vs. a comparable AWS SageMaker solution and as much as 61 percent vs. a comparable Azure ML 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 AI Factory solution, whether they opt to purchase and manage it themselves or engage with Dell APEX Subscriptions. Choosing an on-premises Dell AI Factory 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. CIO, "How to get gen AI spend under control," accessed April 7, 2025, https://www.cio.com/article/3478467/how-to-get-gen-ai-spend-under-control.html.
2. AWS, "AWS Pricing Calculator," accessed April 16, 2025, https://calculator.aws/#/.
3. AWS, "Machine Learning Savings Plans," accessed April 16, 2025, https://aws.amazon.com/savingsplans/ml-pricing/.
4. StackOverflow, "Why should preprocessing be done on CPU rather than GPU?" accessed April 16, 2025, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu.
5. Hugging Face, "Model Memory Requirements," accessed April 16, 2025, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
6. AWS, "Training large language models on Amazon SageMaker: Best practices," accessed April 16, 2025, https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/.

7. AWS, "Machine Learning Savings Plans," accessed April 16, 2025, https://aws.amazon.com/savingsplans/ml-pricing/.
8. 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/.
9. Microsoft, "Azure Pricing Calculator," accessed April 16, 2025, https://azure.microsoft.com/en-us/pricing/calculator/.
10. Microsoft, "Azure Machine Learning pricing," accessed April 16, 2025, https://azure.microsoft.com/en-us/pricing/details/machine-learning/.
11. Dell, "Dell APEX Subscriptions," accessed April 16, 2025, https://www.dell.com/en-us/dt/apex/subscriptions.htm.
12. CrowdStrike, "12 Cloud Security Issues: Risks, Threats, and Challenges," accessed April 16, 2025, https://www.crowdstrike.com/cybersecurity-101/cloud-security/cloud-security-risks-threats-challenges/.
13. CrowdStrike, "12 Cloud Security Issues: Risks, Threats, and Challenges."
14. DataCamp, "The Pros and Cons of Using LLMs in the Cloud Versus Running LLMs Locally," accessed April 16, 2025, https://www.datacamp.com/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally.
15. Meta, "Introducing Meta Llama 3: The most capable openly available LMM to date," accessed April 16, 2025, https://ai.meta.com/blog/meta-llama-3/.
16. Meta, "Introducing Meta Llama 3: The most capable openly available LMM to date."

Read the science behind this report ▶

▶ View the original, English version of this report

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.