The science behind the report:

Accelerate your AI journey while reducing project costs with a validated Dell AI Factory with NVIDIA solution, utilizing Red Hat OpenShift

This document describes what we tested, how we tested, and what we found. To learn how these facts translate into real-world benefits, read the report Accelerate your AI journey while reducing project costs with a validated Dell AI Factory with NVIDIA solution, utilizing Red Hat OpenShift.

We concluded our research on September 23, 2025. The results in this report reflect configurations that we finalized and acquired pricing data for on September 23, 2025 or earlier. Unavoidably, these configurations may not represent the latest versions available when this report appears.

System information

Dell AI Factory on-premises solution

Both payment models of the Dell AI Factory on-premises solution, CAPEX and Dell APEX Infrastructure, include the following hardware:

- 4 x PowerEdge R660 head nodes
- 2 x PowerEdge XE9680 GPU worker nodes
- 3 x NVIDIA SN5600 for network infrastructure
- 1 x NVIDIA SN2201for OOB management

Table 1: Detailed configuration information for each PowerEdge XE9680 GPU worker node.

Configuration information	Dell PowerEdge XE9680 GPU worker node
Number of nodes in solution	2
Chassis	
Chassis	XE9680 6U Chassis with 8 GPU 8x 2.5 NVMe Only
Processor	
Number of processors	2
Vendor and model	Intel Xeon Platinum 8562Y
Core count (per processor)	32 cores and 64 threads
GPU(s)	
Number of GPUs	8-GPU Assembly
Vendor and model	NVIDIA HGX H200 8-GPU SXM 141GB 700W
Memory module(s)	
Total memory in system (GB)	3,072
Number of memory modules	32
Type	RDIMM, 5600MT/s Dual Rank
Size (GB)	96
Local storage	
Number and type of ports	2x 1GbE
Vendor and model	Broadcom 5720 Dual port
Drive size (TB)	3.84
Drive information (speed, interface, type)	Enterprise NVMe™ Read Intensive AG Drive U.2 Gen4 with Carrier
Network adapter 1	
Number and type of ports	2x 1GbE
Vendor and model	Broadcom 5720 Dual port
Network adapter 2	
Number and type of ports	2 x 200GbE
Vendor and model	NVIDIA BlueField-3 Single Port 200GbEQSFP112PCIe Full Height
Network adapter 3	
Number and type of ports	2 x 400 GbE
Vendor and model	NVIDIA BlueField-3 Single Port 400GbEQSFP112PCIe Full Height
Power supplies	
Number of power supplies	1
Vendor and model	3+3 FTR (GPU Power Brake Enabled), Hot-Plug PSU, 2800W MM HLAC (200-240Vac) Titanium, C22 Connector
Wattage of each (W)	2,800

Configuration information	Dell PowerEdge XE9680 GPU worker node
ProSupport and ProDeploy	
ProSupport (5 years)	ProSupport and Next Business Day Onsite Service
ProDeploy Plus	ProDeploy Plus Dell Server XE Series 5U/6U
Embedded system management	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage™ Enterprise Advanced Plus

Table 2: Detailed configuration information for each PowerEdge R660 head node.

Configuration information	Dell PowerEdge R660 head node
Number of nodes in solution	4
Chassis	
Chassis	2.5" chassis with up to 10 hard drives (SAS/SATA), PERC11, 1CPU
Processor	
Number of processors	1
Vendor and model	Intel Xeon Gold 6526Y
Core count (per processor)	16 cores and 32 threads
Memory module(s)	
Total memory in system (GB)	128
Number of memory modules	8
Type	RDIMM 5600 MT/s Single Rank
Size (GB)	16
Storage controller	
Vendor and model	PERC H755 SAS
Local storage	
Number and type of ports	2x 1GbE
Vendor and model	Broadcom 5720 Dual port
Drive size (TB)	3.84
Drive information (speed, interface, type)	Enterprise NVMe™ Read Intensive AG Drive U.2 Gen4 with Carrier
Network adapter 1	
Number and type of ports	2 x 1GbE
Vendor and model	Broadcom 5720 Dual Port 1GbE LOM
Network adapter 2	
Number and type of ports	2 x 100GbE
Vendor and model	Mellanox ConnectX-6 DX Dual Port 100GbE QSFP56 Network Adapter, Low Profile

Configuration information	Dell PowerEdge R660 head node
Cooling fans	
Number of cooling fans	4
Vendor and model	Very High Performance Fan
Power supplies	
Number of power supplies	2
Vendor and model	Dual, Redundant(1+1), Hot-Plug Power Supply,1100W MM(100-240Vac) Titanium
Wattage of each (W)	1,100
ProSupport and ProDeploy	
ProSupport (5 years)	ProSupport and Next Business Day Onsite Service
ProDeploy Plus	ProDeploy Plus PowerEdge R Series 1u2u
Embedded system management	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage Enterprise Advanced Plus

AWS SageMaker solution instances

Table 3: Detailed configuration information for the AWS instances.

Configuration information	ml.t3.medium (notebooks)	ml.r5.16xlarge (processing)	ml.p5en.48xlarge (inference and fine-tuning)
Number instances	20	2	2
Cloud service provider (CSP)	AWS	AWS	AWS
Region	US East (Ohio)	US East (Ohio)	US East (Ohio)
Processor			
Number of vCPU	2	64	192
Memory module(s)			
Total memory in system (GiB)	4	512	2,048
Local storage			
Number of drives	1	1	8
Drive size (GB)	5GB	3500GB	3840GB
Drive information (speed, interface, type)	EBS	EBS	EBS
GPU			
Number of GPUs	N/A	N/A	8
Vendor and Model	N/A	N/A	NVIDIA H100
Additional features	N/A	N/A	3,200 Gbps of networking bandwidth1

How we tested

Introduction

To provide an example for AI solution costs, 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. We sized and estimated the costs for four solutions:

- Dell AI Factory on-premises solution (CAPEX)
- Dell AI Factory on-premises solution managed with Dell APEX Infrastructure
- AWS SageMaker solution

Both payment models of the on-premises Dell solution use the same hardware and software. With the Dell AI Factory on-premises CAPEX solution, the enterprise purchases the hardware upfront; with Dell APEX Infrastructure, Dell installs hardware in the customer's data center and bills the enterprise monthly based on "Committed" and "Buffer Capacity."

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, and we built our scenario around a single, relatively small AI workload. We included costs for data analysts' machine learning development notebooks, data processing tasks, continuous model fine-tuning, and real-time inference.

We sized the hardware for each solution based on assumptions about hours of work and hardware capabilities needed. We used public sources for that research. We used an online pricing calculator for AWS SageMaker and requested and received quotes using Dell Recommended Pricing for the Dell solution for both payment models. We did not do any hands-on testing of any of the solutions for this paper. We did not attempt a 1:1 comparison of NVIDIA AI Enterprise to AWS SageMaker components.

Our findings

In the main report, we show comparisons for how the Dell AI Factory on-premises solution with both payment models (CAPEX and Dell APEX Infrastructure) compared to the AWS SageMaker solution. Tables 4 and 5 show the cost basis for those comparisons. The normalized value is the result of dividing each value by the cost for the Dell AI Factory on-premises solution shown in the table. Table 5 shows that the AWS SageMaker solution would cost up to 1.70x the Dell AI Factory on-premises solution over 5 years

We calculated the pro-rated results using the AWS 3-year commitment prices pro-rated to 5 years. We also calculated the 5-year TCO by pricing three years at the AWS 3-year commitment price and then adding two years each at the AWS 1-year commitment price to reflect the possibility of a customer not opting to re-up for another 3-year commitment. We sourced all prices from the AWS Machine Learning Calculator or regular AWS Pricing Calculator.

Table 4: Normalized 5-year TCO for the Dell traditional on-premises solution compared to AWS SageMaker using pro-rated 3-year subscription pricing and using 3-year commitment plus 1-year commitment pricing.

	Dell AI Factory on-premises solution 5-year costs	AWS SageMaker (pro-rated)	AWS SageMaker (3-year commitment plus 2x 1-year commitment)
Total	$2,121,093.80	$3,429,853.01	$3,605,983.40
Normalized	1	1.61	1.7
Breakeven in months for Dell solution compared to AWS SageMaker	N/A	37.3	35.3

Table 5 shows that the cloud solutions cost up to 1.89x the cost of Dell APEX Infrastructure over 5 years.

Table 5: Normalized 5-year TCO for the Dell APEX Infrastructure solution compared to AWS SageMaker using pro-rated 3-year subscription pricing and using 3-year commitment plus two years of 1-year commitment pricing.

	Dell APEX Infrastructure 5-year costs	AWS SageMaker (pro-rated)	AWS SageMaker (3-year commitment plus 2x 1-year commitment)
Total	$2,295,265.00	$3,429,853.01	$3,605,983.40
Normalized	1	1.49	1.57

Storage considerations

We did not include costs for storage beyond that which is needed for the servers or instances to do their tasks.

The Dell AI Factory on-premises solution with both payment models includes 76.8TB raw storage which, once in RAIDs, would be approximately 38.4TB of usable capacity:

- 15.36TB of raw SSD capacity, or 7.68TB in a RAID 10, on each of the four PowerEdge R660 head nodes
- 7.68TB of raw SSD capacity, or 3.84TB in a RAID 10, on each of the two PowerEdge XE9680 GPU worker nodes

Cluster management and notebook tasks share the storage on the head nodes. Processing, model fine-tuning, and inferencing tasks share the storage on the GPU worker nodes. We provisioned the Dell PowerEdge clusters with some additional storage relative to the cloud solutions to ensure room for management tasks on the head nodes and some room for growth if needed.

The AWS SageMaker solution includes a total of 68.44TB storage:

- 3,500 GB EBS gp2 storage purchased for each of the two ml.r5.16xlarge processing instances for a total of 7,000 GB.
- 8 x 3084 GB NVMe SSDs come default for each of the ml.p5en.48xlarge instances

1 x 5GB EBS temporary storage comes default for each notebook instance included with the instance. Storage needs vary for notebook instances, but typically do not require much for this type of workload, so we opted to leave the cloud instances with the storage that came by default. We included data transfer costs for the EBS data transfer in the AWS solutions. We did not include data transfer costs for the Dell AI Factory on-premises solutions, which would be using on-board SSDs.

Usage hours

We sized the solutions based on the following estimates of hours per month of notebook, data processing, model fine-tuning, and inference usage. We used these hours to calculate hours of instance usage for the cloud solutions and to size those instances and the servers for the on-premises solutions.

We sized the solutions with the assumption that there are 22 workdays in each month, with workloads set to run overnight to maximize usage. Thus, each server and cloud instance would have 528 hours of runtime available each month.

Table 6: Usage hours for the four tasks.

Task	Total hours per month	Usage calculations
Notebook	3,520	We sized each solution to support 20 data professionals, with one notebook instance each, 8 hours a day, 22 days a month, a total of 176 hours each a month, a total of 3,520 for all 20 data professionals. Minimum requirements are small cloud notebook instances with 2vCPU and at least 4GiB memory.
Processing	1,056	Data processing tasks would run during the 528 uptime hours on two Dell PowerEdge XE9680 servers for a total of 1,056 hours runtime and would require 1,056 hours runtime on cloud instances. Minimum requirements for the cloud instances were 64vCPU, 1000GiB memory, 7000 GB storage. We sized the Dell PowerEdge servers to support these requirements plus those of the fine-tuning and inferencing tasks.
Model fine-tuning	792	The combined runtime of 1,056 hours for fine-tuning and inferencing tasks requires two Dell PowerEdge XE9680 servers. All jobs use eight H100 GPUs and up to a half TB of memory. The cloud solutions require the same number of hours on instances with eight H100 GPUs or equivalent. For AWS we used two 8x H100 GPU instances.
Inferencing	264	

Notebooks details

Data scientists would need small cloud notebook instances with 2vCPU and at least 4GiB memory. While some notebook tasks might perform better with more memory, we opted for the AWS ml.t3.medium instance based on the AWS SageMaker TCO guide2 suggestions. For Dell on-premises solutions, we assumed the one core of an equivalent processor and 4.3GB memory per notebook with the notebooks running on the 4x PowerEdge R660 management servers along with management tasks.

AWS SageMaker notebook instances

For the AWS SageMaker solutions, we selected ml.t3.medium notebook instances with 2 vCPUs per instance and 4GiB memory per instance.

Dell AI Factory on-premises solution notebooks

The Dell solution with both payment models supports these notebook workloads on the three PowerEdge R660 head nodes, which combined have 576GB memory and 48 processor cores, enough to support both cluster management tasks and the 20 notebook workloads. Our assumptions for making that sizing decision are as follows:

- Management tasks take up less than half of the processor capacity of these head nodes and less than 500GB memory, with remaining capacity available to run these tasks.
- During the 176 hours each notebook workload runs each month, it would use a single processor core, the equivalent of the 2 vCPU for the SageMaker notebooks, assuming a 1 thread: 1vCPU ratio, and 4.3 GB memory to match the 4GiB we defined in sizing the cloud notebooks. All 20 notebooks running at the same time would use less than half of the 48 cores and 15% of the memory on these systems.
- All work on all systems occurs during less than 72.3% of the total hours in each month, based 22 workdays a week with 24 hours available each day.

Processing details

AWS SageMaker processing instances

Processing tasks run best on CPU rather than GPU and thrive on a high memory to core ratio,[1] so we focused on memory-optimized instances for the AWS SageMaker processing instances. Our target was at least 64vCPU, 1,000GiB memory, and 7,000 GB storage. The AWS SageMaker memory-optimized processing instances didn't offer an instance matching our specification, so instead we selected a pair of ml.r5.16xlarge SageMaker memory-optimized processing instances with a combined 1,024GiB of memory. Together these SageMaker instances exceeded our needs with double the vCPU capacity of our targets.

Table 7: Key configuration information for the AWS SageMaker processing instances.

Instance configuration information	AWS SageMaker ml.r5.16xlarge (processing)
Number instances	2
Number of vCPU	64 (128 for 2 instances)
Total memory in system (GiB)	512 (1,024 for two instances)
Number of drives	1 (2 for two instances)
Drive size (GiB)	3,500 (7,000 for two instances)
Drive information (speed, interface, type)	EBS

Dell AI Factory on-premises solution (CAPEX and Dell APEX Infrastructure)

The two Dell PowerEdge XE9680 worker nodes have more capacity than the fine-tuning and inference workloads require, enough to support the processing workloads running in parallel. The processing workloads would rely on CPUs, and the model fine-tuning and inference on GPUs. To match our target specs, the processing workloads would use at least 1TB of the combined 6 3TB of memory of the two servers, 32CPU cores, and about 3.5TB of the 7.68TB of storage of the two servers.

Model fine-tuning and inference details

The solutions require GPU instances or servers for fine-tuning and inference workloads with eight NVIDIA HGX H200 GPUs 512GB of memory.

AWS SageMaker instances

We chose instances with 8x NVIDIA HGX H200 GPUs and more than our 512GB total memory minimum requirement.

Table 8: Key configuration information for the AWS SageMaker fine-tuning and inference instances.

Inference and training instances	SageMaker ml.p5en.48xlarge
Number of instances	2 (1 for fine-tuning and 1 for inference)
Number of vCPU	192
Total memory in system (GiB)	2,048
Number of drives	8
Drive size (GiB)	3,084
Drive information (speed, interface, type)	NVMe SSD
Number of GPUs	8
Vendor and model	NVIDIA H200

Dell AI Factory on-premises solution

For both payment models of the Dell AI Factory on-premises solution, we sized the two PowerEdge XE9680 GPU worker nodes to handle the fine-tuning and inference workloads using GPU resources and to support the processing workloads using spare CPU and memory resources. The two PowerEdge XE9680 GPU worker nodes each have two 48-core Intel Xeon Platinum 8468 processors, 1,024GB memory, 44.8 TB of raw capacity storage, and an NVIDIA HGX H200 8-GPU assembly.

Cost analysis

For the cloud solutions, we included the licensing cost for the instances we needed for notebooks, processing, fine-tuning, and inference. For the Dell AI Factory on-premises solution, we include two payment options for the hardware: an upfront-purchase model and a Dell APEX Infrastructure monthly payment plan. For both on-premises solutions, we added server administration costs for the hardware and OS, and data center costs for rack space and energy costs for power and cooling, costs that aren't relevant to the two cloud solutions.

We omitted some costs, for example:

- We omitted costs of work that could be similar on all solutions such as installing and maintaining open-source software, data transferring and backup, and the salaries of the data scientists.
- The costs we list for the Dell solutions include the NVIDIA AI Enterprise and Red Hat OpenShift. Otherwise, did not include software costs for any of the solutions. SageMaker instances include some software and services such as Jupyter Notebooks on the notebook instances and processing APIs with the processing instances. We assumed any additional software and tools the data scientists install there would be open source. With the Dell on-premises solutions, data scientists would exclusively use either the NVIDIA and Red Hat software we included in the quote or open-source software and tools such as Jupyter Notebook, Python, and PyTorch.
- We did not include sales taxes because those vary state to state and business to business. We do not include migration costs or end-of-life costs.

We focused on a 5-year lifecycle for each of these generative AI solutions, because we believe it is a reasonable lifecycle for an on-premises generative AI solution that requires state-of-the-art hardware. Organizations could re-purpose the Dell hardware they purchased after that.

Dell solutions

5-year costs for the Dell AI Factory on-premises solution

For the Dell AI Factory on-premises solution with both CAPEX and Dell APEX Infrastructure payment models, we included the following costs over a 5-year period:

- The Dell Recommended Price for Dell servers and switches, including Dell ProDeploy Plus for on-site installation services for the servers, Red Hat OpenShift licenses, NVIDIA AI Enterprise licenses, and a 5-year ProSupport for Infrastructure plan to provide support and maintenance services for the gear.
- System administrator to maintain and secure the hardware and OS
- Energy costs for power and cooling
- Data center costs for rack space

The two solutions included the same hardware and would incur the same costs for system administration, energy costs for power and cooling, and data center rack space costs.

Table 9: 5-year costs for the CAPEX Dell AI Factory on-premises solution.

Dell AI Factory on-premises solution (CAPEX)	5-year costs (rounded up to dollar)
Dell hardware with ProSupport and ProDeploy Plus (for servers), NVIDIA AI Enterprise subscription and support, and OpenShift licensing.	$1,901,350
Server administration	$13,485
Energy costs for power and cooling	$194,884
Data center rack space costs	$11,375
Total	$2,121,094

For the Dell AI Factory on-premises solution, we received a quote using the Dell recommended price. On September 28, 2025, Dell quoted $1,679,521 as the purchase price for the hardware; this price included hardware, Dell support, NVIDIA software, and NVIDIA support. (It did not include the Red Hat OpenShift license, which cost $396,000 for 5 years.) Dell defines recommended price as a price that serves as a starting point for potential buyers. It represents the cost immediately accessible to companies, even if they are not existing customers, and essentially functions as a suggested retail price for their products.

Table 10: 5-year costs for Dell APEX Infrastructure.

Dell APEX Infrastructure	5-year costs (rounded up to dollar)
Dell hardware with ProSupport and ProDeploy Plus (for servers), NVIDIA AI Enterprise subscription and support, and OpenShift licensing.	$2,075,521
Energy costs for power and cooling	$194,884
Data center rack space costs	$11,375
Server administration	$13,485
Total	$2,295,265

Dell AI Factory on-premises solution using Dell APEX Infrastructure

Based on that recommended price, Dell Technologies provided a 5-year cost estimate for Dell APEX Infrastructure of $1,679,571.20. Dell Technologies Sales sent PT a quote for the Dell APEX Infrastructure for the same hardware and software as above for a 60-month term commitment, and a 75% capacity commitment. We received that quote on September 28, 2025. The 75% capacity estimate was the closest capacity option that would cover the 528 uptime hours that we size the solutions to deliver.

System administration

Server administrators monitor and ensure performance, availability, functionality, and security of the hardware and OS, and in this case install the OS. These are services that the on-premises solution requires but the cloud solutions do not because they are included in their service agreement. We estimated a five-year cost of $13,485.00 for this server administration based on the total compensation for a mid-level system administrator[2] who is able to maintain 300 servers and associated switches and OSs using automated tools and processes and who is aided by ProSupport and ProDeploy Plus services.

Dell ProDeploy Plus for Infrastructure and Dell ProSupport for Infrastructure

We did not include a separate estimate for deployment, instead relying on Dell ProDeploy Plus for Infrastructure, a service we included in the hardware quote, to provide onsite hardware and software deployment. A Principled Technologies report shows that ProDeploy Plus for Infrastructure can "Save valuable in-house admin time by using a Dell Technologies-certified engineer for installation and configuration of a Dell solution."

That service might not cover some planning tasks, unboxing and racking the switches, or installing the OS; those tasks would take little time, and are included in the estimate of system administration time to maintain and secure the solution.

We included 5 years of Dell ProSupport and Next Day Onsite Service.

Energy costs for power and cooling

The cloud solutions included energy costs for power and cooling in their prices. For the on-premises solutions, we estimated 5-year power and cooling costs using the Dell Enterprise Infrastructure Planning tool.[3] To get an estimate, we entered in the specifications for the servers and switches included in the Dell Technologies Sales price quote. We provided two other inputs that affected calculations:

- 1.56 power usage effectiveness (PUE) multiplier of power costs to get combined power and cooling costs. That PUE was the industry average in 2023, according to the Uptime Institute, an organization that surveys and tracks data center costs.[4]
- 12.85 cents per kilowatt-hour energy cost based on US Energy Administration (EIA) reported average retail price of electricity for the commercial sector in 2024.[5]

We calculated costs for power and cooling separately for the devices running idle and computational or maximum workloads. We weighed the results based on the 528 runtime hours (about 72.3 percent of an average month) that we sized the solutions to deliver. Note that the Dell tool did not include the NVIDIA switches used in the Dell solutions, so we used the tool to price out the power costs of the PowerEdge R660 servers and PowerEdge XE9680 servers. We then used the power costs of the switches used in a previous Dell TCO study as an estimate of the 4 NVIDIA switches. The actual power costs may differ.

Table 11: 5-year energy costs for power and cooling.

Workloads	Energy cost for 5 years for on-premises solution	Weighting multiplier	Weighted energy cost for power and cooling
5-year energy cost maximum/ computational total	$249,684.57	72.3%	$115,860.42
5 year energy cost idle total	$51,847.64	27.7%	$9,641.28
5-year weighted energy cost			$194,883.74

Data center rack costs

The enterprise would incur additional costs for housing the servers and racks in the data center. These operational expenses (OPEX) include building power and cooling costs, maintenance and repair costs, IT equipment upgrades and replacements, security and staffing costs, and internet and bandwidth costs. We used an online source that stated the average data center is 100,000 sq. ft.[6] and another online source that estimated the operating expenses of a "large datacenter" is between $10 million and $25 million.[7] We used the average of the estimated annual cost rate, $17.5 million, and divided that by the 100,000 sq. ft. data center average to arrive at an estimate of $175 operating expenses annually per sq. ft. of data center space. We estimated that a single standard 24" x 42" rack would need 7 sq. ft. for itself plus an additional 6 sq. ft. behind it for access and exhaust purposes. Thus, we calculated that 4 years of OPEX for one rack would be $2,080.

While we included data transfer for the two cloud solutions to access the AWS S3 storage, we did not add those costs for the on-premises solutions because they would be using their onboard disks or local storage arrays for storage.

AWS SageMaker solution

We configured the AWS SageMaker solution to match the quoted Dell AI Factory on-premises solution as closely as possible for the four tasks we outlined previously: notebooks, processing, model fine-tuning, and inference. The AWS Pricing Calculator for SageMaker lists each task as a separate pricing module you can toggle to add to the estimate. We added SageMaker Studio Notebooks, SageMaker Processing, SageMaker Training, and SageMaker Real-Time Inference. For each module, we filled in the necessary fields to determine the hourly cost of each instance we chose for each task. We then used that hourly cost to determine how much a user would spend to run each instance for the pre-calculated number of hours we determined based on the Dell systems. We calculated twice as many processing instances and therefore processing hours to ensure to match the processing capacity of the other solutions. We also added EBS storage to the processing instances, as they do not spin up with storage outside the OS volume. Because AWS does not offer a 5-year savings plan, we calculated the 5-year TCO two different ways.

First, we assumed that a company would re-up a 3-year plan making the fourth and fifth year cost the same as the first three. While this would require committing to 6 total years, we only present the pro-rated costs of 5 years. Second, we opted to assume our company instead committed to 1 additional year for two years in a row, using the 1-year savings plan. For this scenario, we priced the first 3 years at the 3-year commitment plan price, then priced the fourth and fifth year at the 1-year commitment plan price and added the two together. For the pro-rated scenario, our total costs came to $3,429,853.01 for 5 years. For the 3 + 2 scenario, our total costs came to $3,605,983.40 for 5 years. See Table 12 for the full instance details.

Table 12: 5-year SageMaker solution instance costs.

Service	Instance	Instance $/hr	Run time (hours/mo)	Cost for 5 years (pro-rated)	Cost for 5 years (3+2)
SageMaker Studio Notebooks	ml.t3.medium	$0.02244	3520	$4,739.33	$5,682.12
SageMaker Processing	ml.r5.16xlarge*	$2.2908	2112	$290,290.18	$339,984.69
SageMaker Processing EBS storage (3.5TB per instance per month)	n/a	n/a	n/a	$43,008.00	$43,008.00
SageMaker Training	ml.p5.48xlarge	$48.549355	792	$2,307,065.33	$2,401,185.14
SageMaker Real-Time Inference	ml.p5.48xlarge	$48.549355	264	$769,021.78	$800,395.05
S3 data transfer (1 in and 15 out)	n/a	n/a	n/a	$15,728.40	$15,728.40
Total	n/a	n/a	n/a	$3,429,853.01	$3,605,983.40

1. StackOverflow, "Why should preprocessing be done on CPU rather than GPU?" accessed October 21, 2025, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu and Hugging Face, "Model Memory Requirements," accessed October 21, 2025, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
2. Systems Administrator II total compensation (salary and benefits) of $130,616 per year. Source: Salary.com, "Systems Administrator II," accessed March 25, 2024, https://www.salary.com/tools/salary-calculator/systems-administrator-ii-benefits.
3. Dell, "Dell Enterprise Infrastructure Planning Tool," accessed October 21, 2025, https://dell-ui-eipt.azurewebsites.net/#/.
4. Uptime Institute, "Large data centers are mostly more efficient, analysis confirms," accessed October 21, 2025, https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/.
5. EIA, "Electricity Data Browser," accessed March 29, 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=.
6. Surajdeep Singh, "20 Incredible Data Center Statistics in 2025," accessed October 21, 2025, https://www.hostingadvice.com/how-to/data-center-statistics/.
7. "The Economics of Data Centers: A Deep Dive into Costs and Revenues," accessed October 21, 2025, https://www.concretelogicpodcast.com/blog/the-economics-of-data-centers-a-deep-dive-into-costs-and-revenues/.

Read the report ▶

This project was commissioned by Dell Technologies.

Principled Technologies®

Facts matter.®