Technical Research Report

PROWESS

Financial Services Firms Can Run Digital Marketing Campaigns Faster and Cheaper with On-Premises AI

Tests show how on-premises Dell™ PowerEdge™ servers featuring 4th Generation AMD EPYC™ processors can achieve superior total cost of ownership (TCO) and performance compared to running AI-driven market-analysis workloads in the cloud.

Executive Summary

As financial services enterprises integrate AI into their operations, the demands on their compute infrastructures grow. For many organizations, this growth leads to higher cloud costs. Organizations now realize that cloud solutions might not meet the performance consistency or budget predictability necessary for their crucial workloads. This report, commissioned by Dell Technologies, examines whether on-premises servers—specifically Dell™ PowerEdge™ servers equipped with 4th Generation AMD EPYC™ processors and NVIDIA® GPUs—can deliver a more secure, better-performing, and more cost-efficient alternative to cloud-hosted environments for real-world AI workloads.

Prowess Consulting assessed the total execution time, cost per job, and anticipated total cost of ownership (TCO) of using a hybrid machine learning (ML) and AI pipeline. The pipeline used in this assessment ranges from classic sentiment analysis to innovative applications of transformer-based large language models (LLMs), and it is similar to pipelines used by banking firms for customer analytics and digital marketing. The pipeline mirrors enterprise applications such as customer experience management tools, social media solutions, and semantic search technologies. To conduct this assessment, we compared a Dell PowerEdge R7615 server and a comparable single-GPU instance in Microsoft Azure®. Our analysis measured raw performance and price-performance, while also evaluating operational efficiency, system versatility, and the ability for enterprises to optimize workloads and manage data governance.

Our findings indicate that Dell PowerEdge servers have the potential to surpass cloud-based equivalents in AI workload performance, even with less-powerful GPU resources. Furthermore, our study underscores the architectural strengths of on-premises servers. These strengths include lower latency, consistent resource availability, and tailored hardware configurations that can help avoid the expenses of cloud overprovisioning.

Highlights

Dell™ PowerEdge™ servers enable up to

16% lower

completion time for an AI test workload compared to running the workload on Microsoft Azure®.

Dell™ PowerEdge™ servers provide up to

88% lower

cost per run for an AI test workload compared to running it on Microsoft Azure®.

Dell™ PowerEdge™ servers deliver up to

58% lower

TCO for an AI test workload compared to running it on Microsoft Azure® over a three-year time horizon.

Study Overview

Financial services enterprises, including banks and credit unions, are finding themselves caught between increasing demand for AI-augmented capabilities and the rising operational costs of cloud-based infrastructure. As AI and ML workloads proliferate—particularly in functions like customer experience, analytics, and marketing—many organizations are beginning to question the sustainability of their cloud spend and the consistency of performance they receive from cloud-hosted compute environments. Equally important, many banks and credit unions are expressing growing concerns around data locality, sovereignty, and security. This last concern is especially pertinent when sensitive data must transit third-party cloud environments with limited visibility into where and how that data is processed or stored. At the same time, organizations recognize that access to high-performance infrastructure is no longer optional for modern workloads.

This study, commissioned by Dell Technologies, aims to help financial services businesses explore an alternative to cloud-based AI computing: running AI workloads on modern, GPU-enabled on-premises servers. Specifically, Prowess Consulting tested the hypothesis that Dell PowerEdge servers featuring 4th Gen AMD EPYC processors and NVIDIA GPUs could offer both improved raw performance for AI workloads and superior cost predictability and TCO over time when compared to an equivalent configuration in a public cloud environment.

To ensure relevance to enterprise buyers, the workload we selected for this study is grounded in a real-world marketing analytics use case. Our findings offer insights for financial services organizations seeking a high-efficiency AI pipeline that can be deployed locally as a cost-effective alternative to cloud compute.

Real-World Use Case: Marketing Content Intelligence Pipeline

Modern marketing teams in financial services face a growing crisis of relevance, velocity, and cost. Despite investing heavily in content creation—often involving cross-functional collaboration and external subject-matter experts (SMEs)—marketing organizations frequently struggle to deliver assets that convert. Common pain points include:

- **Speed**: Time-to-market for new content can be unacceptably slow due to lengthy review cycles and data-gathering requirements.
- **Cost**: Content production is expensive, consuming both internal and external resources.
- **Authenticity**: Overreliance on generic AI-generated text can dilute brand credibility and voice.
- **Relevance**: Even compelling content might miss the mark if it does not address real market sentiment or competitive positioning.

These challenges are exacerbated by the fragmentation of marketing channels and the public's growing use of AI-powered search, which can surface outdated or misaligned content and undermine campaign cohesion.

To address these issues, Prowess Consulting developed the proprietary AI-driven pipeline benchmarked in this study. The Prowess LEAP™ workflow ingests a corpus of marketing collateral from internal client sources, clients' competitors, and industry analysts. For this testing, Prowess Consulting focused on performing this market analysis on financial services firms. The rich supply of content in this vertical provided ample material to sufficiently stress all stages of the Prowess LEAP pipeline—particularly CPU-based, ML-driven sentiment analysis at scale and GPU-based embedding of more than 100,000 pieces of content.

The software that powers the Prowess LEAP pipeline is particularly well-suited as a test bed for this study. This pipeline supports marketing teams across industry verticals with AI-augmented content analysis, and it reflects the practical needs of enterprise marketing teams that require deeper insights to inform content strategy. It also makes use of the performance both of classic ML algorithms and of transformer-based embedding processes at the core of the service, which scale proportionally to the performance of the underlying hardware. Specifically, the ML and embedding functions that power the Prowess LEAP service are lightweight enough to run on mid-range hardware, but they readily show performance improvements with high-performance hardware.

Study Methodology

To evaluate the infrastructure demands of this real-world use case, Prowess Consulting implemented the full Prowess LEAP pipeline in both on-premises and cloud environments. The pipeline is divided into five principal phases, each targeting a distinct aspect of the AI workflow:

1. **Topic modeling (CPU-centric)**: Initial topics of interest (in this case, financial-services marketing) are analyzed in a preliminary fashion before serving as a basis for further, deeper analysis.
2. **Sentiment analysis (CPU-centric)**: This phase involves using traditional ML techniques to perform sentiment analysis and extract key topics from diverse marketing collateral; these operations rely heavily on CPU compute, and they are optimized for multithreaded performance and memory access.
3. **Topic embedding (GPU-centric)**: The third phase harnesses AI to provide contextual analysis of the extracted topics; this stage is compute-intensive and benefits from high-throughput GPU acceleration.
4. **Data embedding (GPU-centric)**: For this testing, content from more than 100,000 web pages and other content repositories is embedded at this stage for deep analysis; the heavy compute requirements of this stage benefit greatly from GPU acceleration.
5. **File analysis (CPU-centric)**: The final phase is the actual analysis of the results and uses the CPU.

Together, these phases provide a full-featured enterprise pipeline used to inform and guide marketing decisions. The hybrid nature of the workload—blending classical ML with AI—offers a representative benchmark for organizations seeking to evaluate their servers' readiness for modern AI operations. Moreover, we chose financial services for this testing as an industry vertical with a deep and wide variety of marketing collateral for analysis.

Testbed and Metrics

To ensure a balanced and representative comparison, we executed the full pipeline within two environments configured to deliver comparable hardware capabilities:

- **On-premises environment**: A Dell PowerEdge R7615 server equipped with a single 32-core AMD EPYC 9334 processor and a single double-wide, 300 W NVIDIA® L40S GPU.
- **Cloud environment**: A single-GPU virtual machine (VM) in Azure with the closest available configuration in terms of compute, memory, and GPU capabilities.

Note that the least powerful GPU available for Azure instances powered by 4th Gen AMD EPYC processors is the NVIDIA® H100 GPU. This disparity underscores an advantage of on-premises deployments: the flexibility to right-size hardware for organizational needs (discussed in greater depth in the **Architectural Advantages of On-Premises Dell PowerEdge Servers** section). Table 1 compares core count, memory, and thermal design power (TDP) of the CPUs and GPUs employed in this study as a rough proxy for the power of the respective processors on their own.

Table 1 | Comparative core count, memory, and TDP of the CPUs and GPUs used in this study

Location	Model	Cores/Memory	Thermal Design Power (TDP)	Price[1]
On-premises	AMD EPYC™ 9334	32 cores	210 W	$2,390.15
On-premises	NVIDIA® L40S (AD102GL)	48 GB memory	300 W	$9,971.15
Cloud	AMD EPYC 9V84	96 cores	360 W	Not applicable (N/A)[2]
Cloud	NVIDIA® H100 NVL	94 GB memory	400 W	$33,984.65

For this testing, we collected three key sets of metrics across both environments:

- **Performance**: Total time to execute the complete pipeline from ingestion through analysis.
- **Price-performance**: The cost to execute a single job at scale, normalized for compute-hour pricing and server utilization, in addition to projected power and cooling costs.
- **TCO**: Projected operational costs over a three-year lifecycle, incorporating server acquisition (on-premises) or sustained usage (cloud).

We selected these metrics to reflect both the real-world performance of the workload and the broader financial implications of the infrastructure strategy over time.

Performance Finding: Faster Completion Time On-Premises

In evaluating the performance of AI pipelines across on-premises and cloud compute resources for highly competitive industries like financial services, total execution time remains one of the most business-critical metrics. For the marketing analytics use case at the center of this study, time to insight is a key differentiator—impacting not only the responsiveness of marketing strategies in finance and other industries, but also their relevance in a dynamic marketplace.

Our test results indicate that the on-premises Dell PowerEdge R7615 server delivered an up to 16% lower completion time for the full Prowess LEAP analysis pipeline than the Azure cloud–based configuration (see Figure 1).

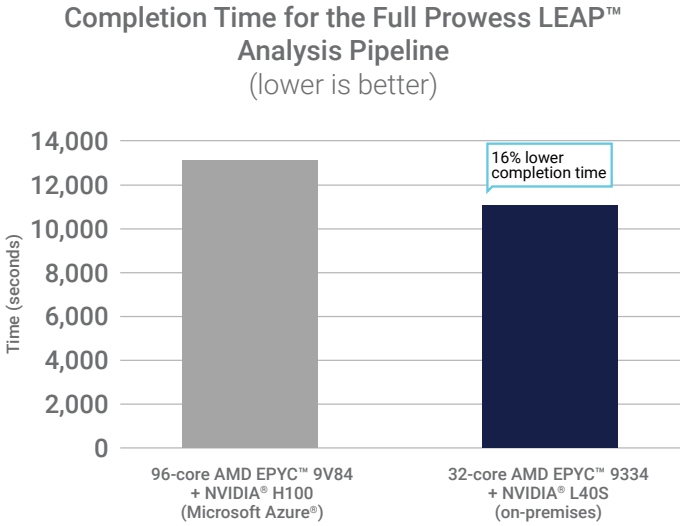

Figure 1 | Completion time for the full Prowess LEAP™ analysis pipeline between a cloud-based Microsoft Azure® instance powered by one 96-core AMD EPYC™ 9V84 CPU and one NVIDIA® H100 GPU versus an on-premises Dell™ PowerEdge™ R7615 server powered by one 32-core AMD EPYC 9334 CPU and one NVIDIA® L40S GPU

This superior on-premises performance comes despite forced overprovisioning in the cloud; cloud instances offer fewer options for CPU and GPU, and thus less granularity for calibrating resources to workloads. In this case, only the AMD EPYC 9V84 CPU with 96 cores paired with an NVIDIA H100 GPU option in Azure could meet or exceed the specifications of the 32-core AMD EPYC 9334 CPU and NVIDIA L40S GPU used in the on-premises Dell PowerEdge R7615 server.

Cost Findings: Price-Performance and TCO Advantage On-Premises

While raw performance is crucial, it is the ratio of performance to cost that often determines server viability at scale. In this study, Prowess Consulting analyzed the cost of completing a single full run of the Prowess LEAP pipeline for both the on-premises Dell PowerEdge server and its cloud-based counterpart.

The on-premises Dell PowerEdge server configuration consistently demonstrated superior price-performance. It completed the same job with an up to 88% lower dollar cost per run of the AI pipeline (see Figure 2). This is especially significant given the forced inclusion of high-end components in the cloud environment, such as the NVIDIA H100 GPU, which typically incur premium pricing and added per-minute charges.

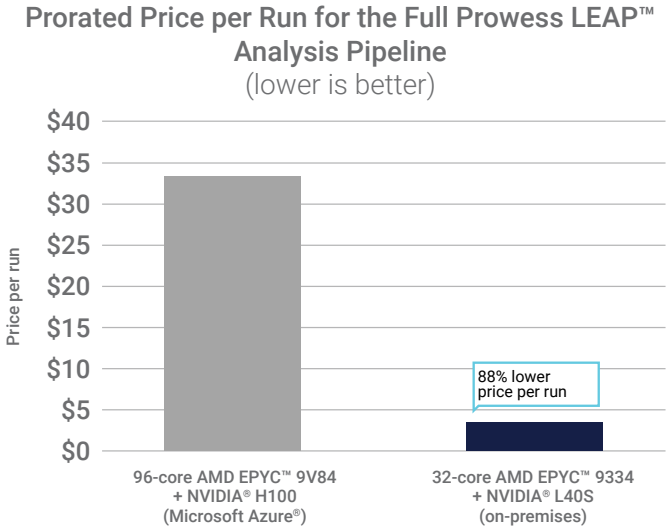

Figure 2 | Prorated price per run for the full Prowess LEAP™ analysis pipeline between a cloud-based Microsoft Azure® instance powered by one 96-core AMD EPYC™ 9V84 CPU and one NVIDIA® H100 GPU versus an on-premises Dell™ PowerEdge™ R7615 server powered by one 32-core AMD EPYC 9334 CPU and one NVIDIA® L40S GPU

These findings suggest that organizations with sustained or predictable AI workloads might see considerable cost advantages by bringing such workloads in-house and on-premises. Moreover, cost savings for AI workloads compound over time. Even accounting for the up-front investment required for on-premises hardware, the Dell PowerEdge server–based configuration yielded an up to 58% lower projected three-year TCO, assuming that the Prowess LEAP workload is only run three times per business day (see Figure 3). (Note that this TCO figure represents the savings even running the on-premises servers at far less than 100% utilization; the closer on-premises utilization grows to 100%, the closer that overall TCO savings will approach the 88% lower cost per individual run.)

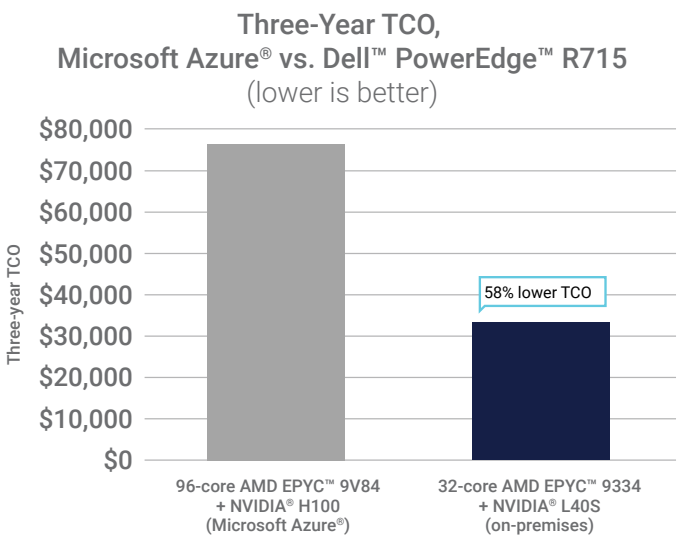

Figure 3 | Three-year TCO assuming running the full Prowess LEAP™ analysis pipeline three times per business day between a cloud-based Microsoft Azure® instance powered by one 96-core AMD EPYC™ 9V84 CPU and one NVIDIA® H100 GPU versus an on-premises Dell™ PowerEdge™ R7615 server powered by one 32-core AMD EPYC 9334 CPU and one NVIDIA® L40S GPU

These results illustrate that, for many organizations, a one-time investment in modern on-premises servers can provide long-term savings—especially for workloads that are run regularly or that require high-performance GPU acceleration.

Financial Predictability

Beyond cost efficiency, on-premises servers offer another key benefit: financial predictability. Fixed hardware and support costs provide a consistent budgeting baseline, eliminating the variability that cloud billing introduces. This is particularly significant with respect to usage-based pricing, reserved instance expiration, and surprise costs tied to resource contention or overage.

This predictability gives financial teams greater control over AI project budgeting, while enabling marketing teams to scale their use of AI-backed tools without fear of unexpected charges. Moreover, savings realized from reduced hardware spend can be reallocated toward higher-value activities, such as additional AI initiatives or expanded content-strategy efforts. In a business landscape that rewards agility, such financial control can be as valuable as the infrastructure itself.

Architectural Advantages of On-Premises Dell PowerEdge Servers

Enterprise infrastructure decisions are often shaped not only by performance and cost considerations, but also by architectural flexibility. In this study, we deliberately configured the on-premises testbed based on a Dell PowerEdge R7615 server with a GPU suitable to the needs of the workload—an NVIDIA L40S GPU—rather than the higher-tier NVIDIA H100 GPU used in the cloud-based test environment.

This decision underscores a critical benefit of on-premises deployments: the ability to right-size hardware to workload requirements. Unlike cloud offerings, which often bundle premium components such as the NVIDIA H100 GPU into fixed instance types—regardless of whether such power is necessary for the target workload—on-premises solutions enable organizations to choose the exact compute components that align with their performance and budgetary goals. In this case, the NVIDIA L40S GPU provided ample acceleration for the Prowess LEAP pipeline's AI needs, without the additional cost or thermal requirements associated with the NVIDIA H100 GPU.

This level of configurability extends to the broader Dell PowerEdge platform, which supports:

- A wide range of current-generation AMD EPYC processors, allowing organizations to match core count and frequency to workload demands.
- Full-bandwidth PCIe® 5.0 support, ensuring high-throughput connectivity to GPUs and storage.
- High-capacity DDR5 memory, improving data-access speeds and system responsiveness for memory-intensive AI tasks.
- Multiple GPU types and models (including AMD, Intel®, and NVIDIA GPUs) and multiple GPUs (specifically, the Dell PowerEdge R7615 server supports up to three double-wide or six single-wide GPUs).

In stark contrast, enterprises relying on cloud service providers (CSPs) can often find themselves constrained by the fixed configurations and instance types offered. CSPs often package premium components into standard instance types, which might not always correspond with precise workload needs or financial considerations.

On-premises server hardware provides organizations with an infrastructure that they can tailor to their own priorities—whether those are cost-efficiency, performance, power optimization, or a combination thereof. Additionally, on-premises Dell PowerEdge servers offer environmental and operational efficiency benefits. With compact 1U and 2U form factors supporting high-core-count CPUs and double-width GPUs, Dell PowerEdge systems deliver full-stack AI performance in a space-saving configuration. These efficiencies are particularly valuable for organizations seeking to scale their AI infrastructures without expanding rack space or increasing heating, ventilation, and air conditioning (HVAC) loads. Moreover, Dell Technologies complements this hardware efficiency with a mature suite of systems-management tools that reduce overhead and optimize daily operations.

Workload Control and Optimization

Another core advantage of on-premises deployment is the full control it offers over the software stack, BIOS settings, and resource pinning for predictable performance. With Dell PowerEdge servers, organizations can fine-tune BIOS settings, implement custom software stack configurations, and pin workloads to specific cores or non-uniform memory access (NUMA) nodes for optimal performance. These levels of control are often unavailable—or at best limited—in cloud environments, where abstraction layers and shared tenancy can prevent low-level system tuning.

Furthermore, on-premises deployment completely eliminates the "noisy neighbor" effects and startup latency common in shared cloud environments. In a shared cloud instance, contention for CPU or GPU resources can result in unpredictable performance—even when reserved instances are used. With the on-premises Dell PowerEdge platform, physical isolation ensures deterministic execution patterns, which is especially critical for ML workloads that rely on consistent processing times for benchmarking and model validation.

While virtualization remains a valuable tool for many organizations—particularly for efficient resource allocation and streamlined management—on-premises deployments offer direct hardware access that can maximize performance for demanding workloads. For AI and ML systems where predictable, high-throughput computing is critical, the absence of virtualization overhead means that bare-metal servers can deliver more consistent and optimized results than cloud VMs. Support for NUMA-aware tuning and core-level optimization is particularly valuable for mixed CPU/GPU pipelines, allowing for more efficient resource utilization and faster processing times.

Finally, on-premises Dell PowerEdge systems ensure that GPUs are always available on demand, with no queueing or provisioning delays. Startup latency is a common friction point in using cloud deployments. Launching GPU-backed instances often involves queue times, cold starts, or provisioning delays. In contrast, the on-premises Dell PowerEdge system remains continuously available, with all resources ready for immediate use, shortening response times and allowing for real-time analysis when needed.

Security and Data Governance

On-premises servers offer inherent advantages in data protection, regulatory compliance, and governance. By keeping sensitive or proprietary datasets entirely within the enterprise perimeter, organizations reduce their exposure to third-party risk, mitigate compliance burdens, and maintain greater control over access policies and audit trails.

In the financial services sector, regulatory standards such as the Gramm-Leach-Bliley Act (GLBA) are critical. By leveraging on-premises servers, financial services organizations can ensure that their data handling practices comply with stringent standards like GLBA, which mandates robust measures for protecting consumer information. Supporting fine-grained identity and access control through integration with enterprise-wide identity and access management (IAM) systems helps meet these regulatory requirements. This ensures that only authorized personnel have access to critical data, and access can be closely monitored and audited.

Furthermore, on-premises deployment helps avoid the regulatory complexity associated with cross-jurisdictional data transfers. In cloud environments, securing datasets often involves navigating complex configurations, cross-region access controls, and coordination with third-party providers. With on-premises servers, these risks can be addressed directly through network segmentation, physical access controls, and comprehensive IAM policies, simplifying compliance with data protection regulations.

Dell Technologies Systems Management

Dell Technologies provides a comprehensive suite of systems management tools designed to streamline IT operations, optimize resource utilization, and enhance overall server efficiency. At the heart of this suite is Integrated Dell™ Remote Access Controller (iDRAC), which offers agentless, out-of-band server monitoring and recovery. This management tool enables administrators to oversee server health, update firmware, and troubleshoot issues remotely, ensuring minimal disruption to operations and reducing the need for on-site intervention.

Complementing iDRAC is Dell™ OpenManage™ Enterprise, which extends iDRAC's capabilities across the environment. Dell OpenManage Enterprise provides a centralized, policy-based management solution for server fleets. It enables IT teams to automate and standardize server-management tasks across their entire environments, providing a cohesive and efficient approach to infrastructure oversight. This centralization facilitates consistent policy enforcement and simplifies the administration of large-scale deployments.

Further enhancing this management ecosystem are add-on modules such as Dell™ Power Manager, which deliver granular insights into energy and thermal usage. These modules enable organizations to comply with internal sustainability and energy efficiency targets by monitoring and adjusting power consumption and thermal profiles in real time. The ability to enforce energy-saving policies at the server-, rack-, or facility-level helps organizations minimize their environmental footprints while maintaining optimal performance.

Unlike public cloud environments, in which management capabilities are often abstracted and restricted to the service provider's interface, on-premises Dell Technologies systems' management tools offer both depth of functionality and customization. In public cloud deployments, administrators are typically limited to the configurations and controls provided

by the CSP, which can hinder the organization's ability to fine-tune resources for specific workloads. On-premises Dell Technologies systems, however, grant full access to BIOS settings, firmware updates, and hardware configurations, enabling a level of control and optimization that is not achievable in the cloud.

Moreover, the direct access to hardware provided by Dell Technologies management tools eliminates the latency and performance variability commonly encountered in shared cloud environments. A prime example of this in public cloud instances is the noisy neighbor effect and the unpredictable performance that can result from contention for shared resources. With on-premises Dell Technologies solutions, resources are dedicated and isolated, ensuring consistent and reliable performance, which is crucial for demanding applications such as AI and ML workloads.

Long-Term Efficiency Benefits

Dell PowerEdge platforms support a range of technologies that promote long-term operational value:

- EPEAT® Silver certification, indicating compliance with stringent environmental criteria
- Titanium-grade power supplies, maximizing power-conversion efficiency to reduce energy waste
- Advanced cooling technologies, including Multi-Vector Cooling 2.0 and Dell™ Smart Flow airflow optimization
- Direct Liquid Cooling (DLC) options for thermally demanding deployments, reducing cooling energy usage

These innovations support high-performance workloads, such as sustained GPU-based AI pipelines, while minimizing their environmental costs. By operating more efficiently, Dell PowerEdge servers also enable server scaling without proportionate increases in energy or thermal load—an important consideration for environmental, social, and governance (ESG)-conscious organizations.

Moreover, these efficiency and sustainability benefits are superior to those offered by CSPs. CSPs often do not offer the same level of granularity in power management or cooling technologies, making it harder to achieve the same degree of operational efficiency and environmental responsibility. With on-premises Dell Technologies solutions, organizations have direct control over their hardware, allowing for precise optimization that can significantly reduce energy consumption and enhance sustainability. This distinction is crucial for enterprises aiming to meet rigorous ESG goals while maintaining robust performance.

Conclusion

Prowess Consulting evaluated the cost and performance tradeoffs between cloud-based and on-premises servers for a real-world AI/ML workload. Our aim was to assess cloud versus on-premises solutions for our own production workload. We found that the on-premises solution was faster in raw execution, and the Dell Technologies platform cut the cost per job by up to 88% and offered substantial TCO advantages over a three-year horizon. This outcome was especially significant given that the cloud deployment used higher-end components, making it a less efficient and more expensive solution overall.

In real-world terms, the up to 16% lower completion time afforded by on-premises Dell Technologies servers can translate to financial services firms being able to run their AI pipelines approximately seven times in the time it would take a competitor to run the same pipeline only six times using a comparable cloud solution. Performance benefits add up over time, particularly in hyper-competitive industries like finance.

The results also reflect architectural alignment between Dell Technologies servers and the demands of modern, production-grade AI workloads. On-premises deployments enable fine-tuned workload optimization, remove reliance on shared resources, and support predictable budgeting. These benefits matter most for organizations with ongoing AI needs—not one-off experiments—particularly when data locality, compliance, or reproducibility are at stake.

Our test results point to the Dell PowerEdge platform as a serious contender for financial services enterprises reassessing their cloud commitments in light of AI scale-out challenges. IT organizations should consider on-premises servers not as a fallback or legacy option, but as a forward-looking, cost-efficient strategy for AI acceleration. For financial services firms, leveraging on-premises Dell Technologies servers can mean achieving faster processing times and considerable cost savings, which are crucial for maintaining a competitive edge and ensuring robust, reliable performance for critical AI applications.

To learn more about Dell PowerEdge servers with 4th Gen AMD EPYC processors, visit: **dell.com/en-us/dt/servers/amd.htm**

For additional research and insights from Prowess Consulting, visit: **prowessconsulting.com/labs**

Appendix A: Hardware and Cloud-Instance Specifications

Table 2 | System specifications

	Standard NC40ads H100 v5 (Microsoft Azure®)	Dell™ PowerEdge™ R7615 (On-Premises)
CPU	96-core AMD EPYC™ 9V84 processor	32-core AMD EPYC™ 9334 processor
Number of CPUs	1	1
Cores/threads per CPU	40/40 (number of vCPUs)	32/64
Cores/threads total	40/40 (number of vCPUs)	32/64
CPU frequency	2.4 GHz	2.7 GHz
Installed memory	320 GB	384 GB
Memory DIMM	N/A	16,384 MB SK hynix® semiconductor DDR5
Memory speed	N/A	4,800 MHz/4,000 megatransfers per second (MT/s)
Number of memory DIMMs	N/A	12
Graphics card	NVIDIA® H100 NVL	NVIDIA® L40S AD102GL

Appendix B: Test Results

Table 3 | Test-run times, by stage and total (median value of three test runs)

Stage	Standard NC40ads H100 v5 (Microsoft Azure®)	Dell™ PowerEdge™ R7615 (On-Premises)
Topic modeling (CPU-focused)	24.8 s	34.2 s
Sentiment analysis (CPU-focused)	18.7 s	24.7 s
Prompt embedding (CPU-focused)	10.2 s	12.2 s
Data embedding (GPU-focused)	13,207.0 s	10,983.0 s
File analysis (CPU-focused)	2.0 s	3.6 s
Total run time	**13,262.7 s**	**11,057.7 s**

Table 4 | Initial system and cloud instance cost per hour (includes instance and storage cost for Microsoft Azure®)

Cost	Standard NC40ads H100 v5 (Microsoft Azure®)	Dell™ PowerEdge™ R7615 (On-Premises)
System capital expenditure (CapEx) cost	N/A	$30,777.35
System CapEx cost per hour (prorated over three years)	N/A	$1.17
Instance cost per hour	$9.18	N/A

Table 5 | Maximum processor power consumption and energy consumption per test run on premises

Processor	Processor Maximum Power Consumption	Projected Maximum Energy Consumption per Run[3]
AMD EPYC™ 9334 (on-premises)	240 W	0.01 kWh
NVIDIA® L40S (on-premises)	350 W	2.14 kWh

Table 6 | Projected cost of energy and cooling per test run (on premises)

	Average Price of Commercial Electricity (United States)[4]	Projected Cost of Energy and Cooling per Run
Dell™ PowerEdge™ R7615 (on-premises)	$0.1309/kWh	$0.28

Table 7 | Per-run and total three-year projected cost, on premises and in the cloud

Total Cost	Standard NC40ads H100 v5 (Microsoft Azure®)	Dell™ PowerEdge™ R7615 (On-Premises)
Per run (prorated per second)	$33.83	$3.87
Over three years (assumes three runs per day)	$76,117.50	$30,777.35

Endnotes

[1] All pricing supplied by Dell Technologies and is current as of August 2025.

[2] Because the AMD EPYC™ 9V84 is a cloud-only CPU SKU, pricing for it is unavailable.

[3] We doubled the maximum power use for the AMD EPYC™ 9334 processor and the NVIDIA® L40S GPU to account for power draw for cooling and for other system components (such as storage and fans).

[4] U.S. Energy Information Administration. "**Average Price of Electricity to Ultimate Customers by End-Use Sector**." Accessed June 2025.

0825/250013