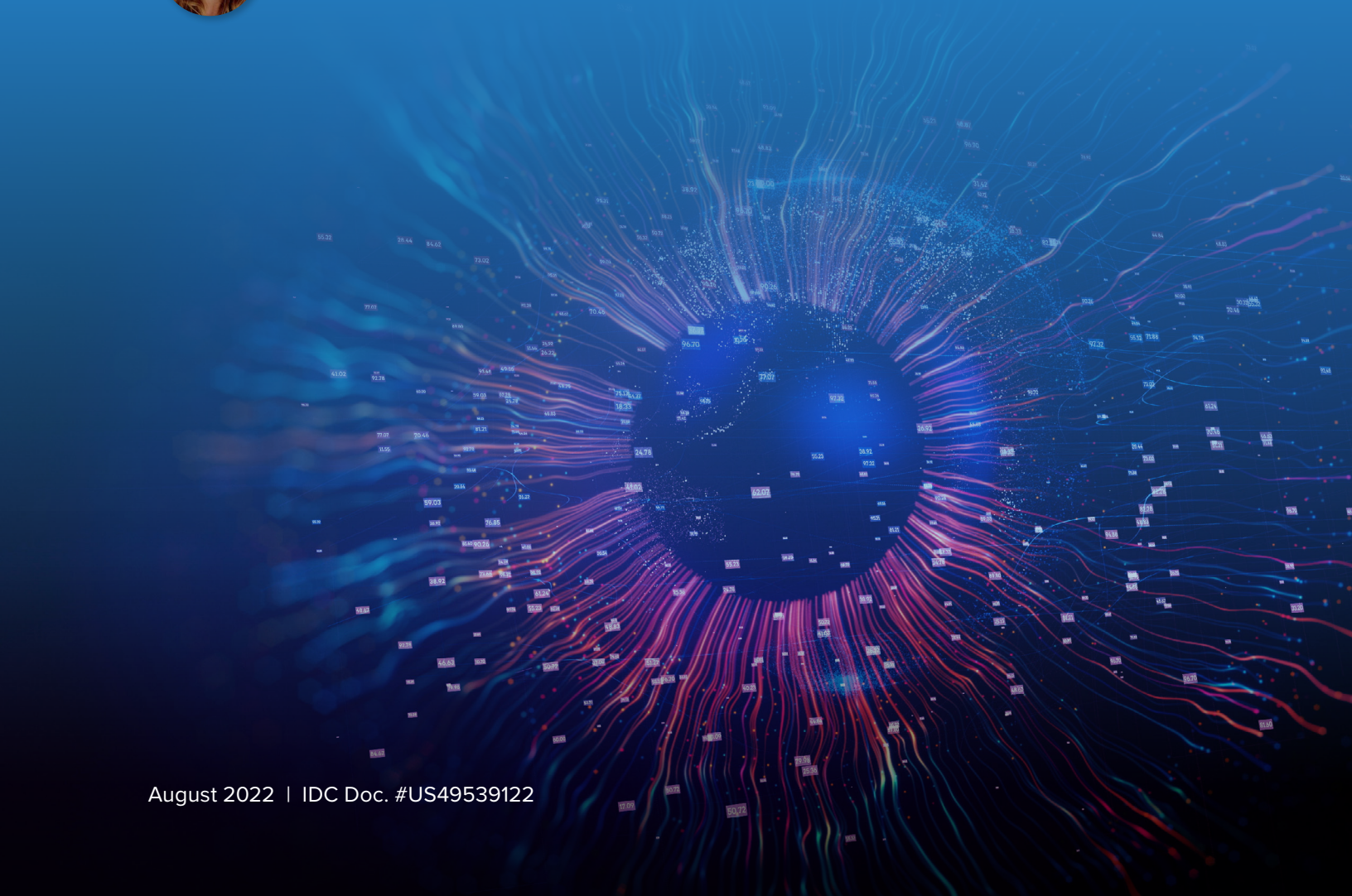
IDC

What Businesses with AI in Production Can Teach Those Lagging Behind

RESEARCH BY:

Peter Rutten
Research Vice-President, Infrastructure Systems, Platforms and Technologies Group, Performance-Intensive Computing Solutions Global Research Lead, IDC

Penny Madsen
Senior Research Director, IDC BuyerView Research, IDC



August 2022 | IDC Doc. #US49539122

Table of Contents

Click on titles or page numbers to navigate to each section.

IDC Opinion

Businesses are in different stages with their artificial intelligence (AI) technologies. Even today, after many years of evaluating AI, only one third of organizations have actually put AI into production (31%[1]). Organizations that started somewhat later are currently prototyping their AI solutions (20%). Companies that only recently came on board with AI are experimenting with AI technologies for their business cases (25%). And a relatively large portion (24%) have only just begun evaluating AI for their business cases.

The urgency to start using AI differs by industry, by company size, and by company strategy. But IDC believes that we have now reached a stage at which every organization must have a solid approach to incorporating AI into its processes and/or offerings to remain viable in the coming years. This means that organizations in the last group — those that are still evaluating — are in trouble in terms of their capabilities and their ability to compete.

What can this group do to catch up faster? Are they doomed to make the same mistakes the front-runners made in the last several years as they ramped up their AI stance? IDC believes that this can be avoided if those that are lagging (we will call them "AI evaluators") inform themselves about what those that are ahead and in production (we will call them "AI users") are doing today.

The purpose of this paper is to compare what those in production with AI do differently compared with those still evaluating, so as to support the latter with these best practices as they move forward on their AI journey. It will also very briefly discuss a few Dell solutions that can support organizations' AI initiatives.

1 All data in this white paper is from IDC's AI InfrastructureView. For more information about this study, see the section "Methodology" at the end of the document.

Situation Overview

The More Sophisticated the AI Model, the More Potential It Has for the Business

AI Users Are Further Ahead in Building Complex AI Models than AI Evaluators

The larger and more complex an AI model is, the greater its potential to accurately predict a trend, recommend a purchase, translate a sentence, or recognize an image. Larger and more complex AI models require greater skills and more sophisticated hardware and software.

AI users are further ahead than AI evaluators in building AI models of various complexity levels, as follows:

33.3% more

AI users than AI evaluators are in the stage of using content analytics, discovery, search, text mining, cognitive platforms, rich media analytics, natural language processing, and regression.

25.5% more

AI users than AI evaluators are building single-layer models that devise algorithms for predictions based on training data, using such techniques as supervised learning, dimension reduction, or outlier detection.

9.8% more

AI users than AI evaluators are in the stage of developing multi-layered machine learning; this includes using techniques such as recurrent, recursive, and convolutional neural networks as well as unsupervised pretrained networks.

More sophisticated AI models enable AI users to gain more impactful and — importantly — more accurate outcomes from the AI functionality they build into their applications. Accuracy translates directly into end-user satisfaction. If the end user is, for example, a line-of-business worker, their productivity will increase. If the end user is a customer, their appreciation of a service or product will increase. In both instances, the ultimate business benefits can be measured in revenue improvements.

Note that the AI users are not much further ahead with the most complex models than the AI evaluators. This is because the evaluators have an opportunity to leapfrog the simpler techniques that AI users started out with. Rather than go through every step of the AI model complexity evolution, AI evaluators can go straight to the more complex models. But they will need to know what will be required to take those models into production.

Challenges with AI Are Common

AI Users Experience Fewer Challenges than AI Evaluators

Developing an AI solution is challenging, but practice makes it easier: 26.8% more AI users than AI evaluators said they do not expect any barriers with their continuing AI work. In other words, developing AI gets more straightforward the further along the organization is. For example, fewer AI users (12.8%) than AI evaluators see any barriers caused by not having the right data sets with which to develop their AI models. They also have less trouble attracting data scientists and on average spend less time preparing a completed machine learning (ML)/deep learning (DL) model for deployment.

But AI users do have more mature challenges to a greater extent than AI evaluators, specifically with:

- Bringing in MLOps expertise — the skills of managing the entire lifecycle of AI development and deployment
- Ethical questions around AI: Are the AI models biased? Could they be used unlawfully or unethically targeting certain demographics?
- Data privacy issues with data used for AI training, as well as with the data that is generated by an AI application

Between 20–30% more AI users experience such challenges than AI evaluators. For the latter, this provides guidance to start addressing these issues from the start. Doing so will prevent much disruption and accelerate results when AI models are in production.

AI Infrastructure Is Critical

AI Users Do a Better Job of Using AI Infrastructure than Evaluators

Servers, storage, and networking for AI often have more advanced components than general-purpose hardware. Most AI evaluators are already aware of this, having laid out, on average, $15.8 million annually for their AI hardware. AI users spend even more, though ($17.6 million), and they expect to spend more in the future than evaluators do. What AI users do with this larger budget is optimize AI results. As a result, IDC data shows, AI users have fewer problems with processor performance, co-processor performance, storage

I/O, and infrastructure management than AI evaluators do. Most importantly, AI users have integrated their AI platforms with the rest of the datacenter and the cloud, whereas AI evaluators are still operating on AI siloes. Almost 49% more AI evaluators than AI users say that they run AI in silos that are used by separate groups without a formal organization-wide strategy or without any coordination as part of a broader vision.

This is a critical deficiency that AI evaluators must address to maximize success. By overcoming performance problems, they will iterate faster on AI models during the training phase, allowing data scientists and AI engineers to build more sophisticated models, do more iterations for improved accuracy, or use the time gained to develop a greater diversity of AI models. The difference can be as significant as several weeks less spent on iterating. Once in production, performance matters for enabling near-real-time AI inferencing, a key requirement for successful AI functionality.

AI Users Focus More on Performance, Scalability, Data Integration, and Cost

AI users have different priorities than AI evaluators for their AI infrastructure, and the latter should be more focused on adopting those same priorities.

For example:

27.9% more AI users than AI evaluators say that performance for AI model training is a key infrastructure requirement; AI evaluators tend to remain on less-performant hardware for too long, potentially failing to see the benefits AI users already enjoy.

14.3% more AI users than AI evaluators say that scalability from AI model development to production is an important requirement; they are further ahead and need to scale their infrastructure for anticipated demand when an AI model goes into production. But AI evaluators should plan well for this stage.

22.8% more AI users than AI evaluators follow their IT team's recommendations for selecting AI compute infrastructure on premises. IT recommendations are the top driver for AI infrastructure acquisitions, well above data scientist, developer, or vendor recommendations; AI evaluators will benefit from partnering with their IT organization on any AI initiative and heeding the team's advice.

14.3% more AI users than AI evaluators are making sure that ease of integration with data repositories is a key AI storage requirement. AI storage infrastructure had been evolving more slowly than compute but is now catching up, and AI evaluators have an opportunity to leverage new AI storage solutions that did not exist a few years ago.

As such, AI users are also more focused on the long term, with **46.8% more** AI users than AI evaluators purchasing on-premises storage solutions with an eye on the long-term costs. AI evaluators need to plan for dedicated storage for their AI initiatives, with a clearly calculated ROI.

AI Users Understand the Need for GPU Acceleration

The most prevalent technology today for boosting AI performance is GPU acceleration. This is widely understood among both AI users and AI evaluators; both use GPUs extensively for AI training and AI inferencing. But AI users have extended this acceleration technology more widely than AI evaluators. One area in which AI users report acceleration benefits is in edge deployments: 13% more AI users than AI evaluators use edge-based GPU acceleration for training and inferencing. AI evaluators will discover that placing AI solutions closer to where data is created will become critical to enabling success.

Even as GPUs are a well-understood technology by both AI users and AI evaluators, the big difference between the two groups is in the number of GPUs they use: 57.4% more AI users than AI evaluators run their AI workloads on (on average) six GPUs per server, 40.9% more AI users than AI evaluators are planning to run these workloads in the near future on seven GPUs per server, and 69.3% more AI users than AI evaluators are moving to running them on servers with (on average) eight GPUs. These are exceptionally large performance differences that AI evaluators will need to catch up on in order to reach production at scale for their AI workloads.

AI Users Use Specific Storage for AI

As mentioned above, storage for AI is becoming increasingly purpose-built for optimizing analysis. The I/O demands of AI and the parallelization with which AI is processed have encouraged storage vendors to design AI-specific storage solutions with parallel file systems that perform much better than general-purpose storage. Organizations are increasingly using these solutions, and AI users are doing so more widely than AI evaluators.

For example:

13.5% fewer

AI users than AI evaluators use their AI storage for workloads other than AI, with AI users saying that the reason for this is that they do not want to compromise their AI storage performance and scaling.

22.8% more

AI users than AI evaluators use commercial storage systems that run parallel file systems.

19.0% more

AI users than AI evaluators use unified data lakes or data platforms that support structured and unstructured data in a single repository for AI.

For AI evaluators, the message is to plan for all aspects of the infrastructure to maximize returns.

There Are Various Scenarios for AI Deployment

Both AI Users and AI Evaluators Fully Leverage Their On-Premises Environment

AI is a distinctly different workload than all others, with a long lifecycle that consists of multiple stages, each with its own infrastructure requirements. Not unlike for high-performance computing, performance and meeting service-level agreements (SLAs) are core metrics for success in AI development and deployment. Data scientists want to know how fast they can iterate on an AI model given its complexity, number of parameters, and accuracy requirements. Line-of-business managers want to know how fast a model can scale and be inferenced for external demand.

One organization IDC spoke with on this subject stated, "One of our prime motivations for having our own datacenter is that a significant part of our research and development work is done in-house." They also said, "We are able to significantly reduce costs by having our own colocated space, even though it's a fixed set of resources. … In addition to saving over 50% in both capex and amortized capex and opex, we're actually able to improve our performance and throughput of our models by two to three times on average. We can also read and write data … substantially faster than we could do from a different location."

Because of these considerations, AI users today run 43% of their AI environment on premises, even as they have experimented with cloud deployments. In some cases, they have moved AI workloads from the cloud back to on premises. AI evaluators also run a large portion of their AI workloads on premises (46%), but for them this is because they started out on premises, not because they tried alternatives. AI evaluators should therefore keep in mind that taking AI into production successfully will mean continuing to run a significant portion of their AI on premises, even as they move other workloads to the cloud.

AI Users Leverage On Premises (and Edge) Much Better for Specific Algorithms

If we take a deeper look at how certain types of AI algorithms and different deployment scenarios are related, it turns out that AI users are specifically focused on managing their own infrastructure. For advanced analytics and statistics, for example, both AI users and AI evaluators run 51% of these workloads on premises today, but 28.3% more AI users than AI evaluators plan to do so in the near future. What's more, 46% more AI users than AI evaluators run these algorithms at the edge.

The on-premises portion of algorithms for machine learning is even larger (59% for AI users and 61% for AI evaluators), but 35.6% more AI users than AI evaluators will run ML algorithms at the edge. Similarly, AI deep learning is a significant on-premises workload (48% for AI users and 70% for AI evaluators), yet here too, 41.4% more AI users than AI evaluators will run deep learning algorithms at the edge in the near future.

What AI evaluators can learn from this is that while AI workloads can be run in many locations, optimizing ROI requires control of the AI infrastructure put in place.

About Dell's AI Solutions

AI users see many benefits from their proactive stance on AI technology. However, Dell Technologies believes it is not too late for AI evaluators or organizations that are further ahead than evaluators but that are not yet in production with AI.

Dell has specific solutions to help these businesses accelerate their AI adoption.

- AI users are leveraging their expertise and infrastructure in a wide variety of areas. With the Dell Validated Design for AI, organizations that want to catch up can quickly build AI architectures for a wide variety of solutions, including deep learning, machine learning, and machine learning operations. They should focus on developing and deploying a comprehensive AI infrastructure strategy, while simplifying their IT to provide faster, deeper insights. Dell's server, storage, and hyperconverged infrastructure (HCI) portfolios are specifically designed for this purpose.
- AI users also focus on performance, scale, and cost and have invested in specialized hardware to enable that. With the Dell Validated Solution for AI, organizations can run many AI workloads in their existing VMware environment. They can get a head start on AI projects without requiring specialized ops expertise, while maximizing their existing investments. Dell solutions are designed to power any AI workload with highly performant, scalable, and easy-to-manage IT infrastructure that supports on-premises, cloud, edge, or hybrid environments, with protection across boundaries.
- Many AI users have built their AI initiatives on on-premises GPU acceleration technology. Dell's partnership with NVIDIA and its portfolio of high-performance GPUs and related solutions is supporting organizations across industries to deliver on their AI initiatives faster and with greater accuracy. For many of these businesses, this means they achieve a continuous stream of real-time insights at scale that is essential for their business goals.

With its solutions, Dell Technologies aims to support small and large organizations in accelerating their AI initiatives, moving to production faster, and becoming successful AI users.

Challenges/Opportunities

For Organizations

Businesses that are still evaluating AI initiatives have to worry about catching up, but at the same time they have the advantage of being able to leapfrog. They can learn from the mistakes early adopters (who were essentially guinea pigs) made and avoid them. This means they will be able to move faster than the front-runners did at the time, taking advantage of new solutions that did not exist a few years ago as well as benefiting from a general consensus as to how AI is best developed and deployed. The most important strategy for exploiting this advantage is a willingness to learn and to not attempt to reinvent the wheel.

For Dell

As a prominent server and storage vendor in the AI infrastructure market, Dell has helped many businesses become AI-literate. However, as in any market, some organizations were left behind during this technology revolution. Businesses that saw no reason to get into AI dropped further and further off the radar, until technology vendors realized that these organizations need help, lest they become obsolete. For Dell, this means demonstrating to the AI evaluators that today, it is much less complicated to get a working AI solution off the ground than it was five or even three years ago. All Dell needs to do is point at the successes that have been achieved by those in production with AI today and help their customer leapfrog to a similarly advanced stage in their AI development.

Conclusion

IDC research has found that businesses that are in production with AI have the following differentiating characteristics from those that are still evaluating AI:

1. They focus on performance, scalability, data integration, and cost.
2. They leverage AI-specific infrastructure with much higher GPU attach rates.
3. They maintain a solid on-premises and edge deployment, and match specific AI algorithms to run on premises.

As a result, AI users run more sophisticated AI models in production that deliver important new functionality to their products and processes, all of which translates directly into improved customer satisfaction and business revenue.

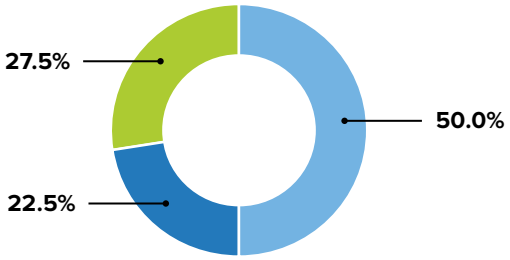
IDC also believes that Dell's AI solutions portfolio is particularly suitable for helping AI evaluators speed up their AI development and deployment, catch up with the early adopters, and improve their competitiveness as a result.

Methodology

The data in this white paper is from IDC's AI InfrastructureView, a deep-dive benchmarking study on infrastructure and infrastructure-as-a-service adoption trends for AI/ML use cases.

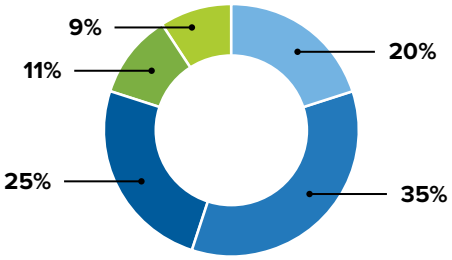
The study consists of a survey among 2,000 IT and line-of-business executives who influence their organization's purchasing of AI infrastructure deployments, services, systems, platforms, and technologies.

Region



- North America
- EMEA
- Asia/Pacific (including China and India)

Organization size (employees)



- 100 or less
- 100 to 999
- 1,000 to 4,999
- 5,000 to 9,999
- 10,000 or more

All industries are represented, with information technology, healthcare, retail, banking, education, manufacturing, government, and professional services being the largest.

About the Analyst

Peter Rutten

Research Vice-President, Infrastructure Systems, Platforms and Technologies Group, Performance-Intensive Computing Solutions Global Research Lead, IDC

Peter Rutten is a Research Vice-President within IDC's Worldwide Infrastructure Practice, covering research on computing platforms. Mr. Rutten is IDC's global research lead on performance-intensive computing solutions and use cases. This includes research on High-Performance Computing (HPC), Artificial Intelligence (AI), and Big Data and Analytics (BDA) infrastructure and associated solution stacks. In this role, Mr. Rutten leads three IDC programs: High-Performance Computing Trends and Strategies, High-Performance Computing as a Service, and Infrastructure Trends and Strategies: Artificial Intelligence and Analytics. His coverage of performance-intensive computing includes supercomputing as well as institutional and mainstream high-performance computing, high-end, accelerated, in-memory and heterogeneous computing infrastructure systems, platforms, and technologies. It includes computing platforms with GPUs, FPGAs, ASICs, and other accelerators that are deployed in the cloud as well as on-premises. It also includes research on mission-critical x86 platforms, mainframes, and RISC-based systems as well as their operating environments (Linux, z/OS, Unix). Mr. Rutten also examines emerging technologies and platforms such as quantum computing, neuromorphic computing and technologies that are potentially disruptive to mature infrastructure markets. As part of his role, Mr. Rutten performs quantitative (market sizing and forecasting) and qualitative (primary research based) analysis as well as custom market sizing for IDC's clients.

More about Peter Rutten

Penny Madsen

Senior Research Director, IDC BuyerView Research, IDC

Penny Madsen is a senior research director for IDC's BuyerView research, focusing on Cloud Pulse, which provides quarterly insights into cloud adoption and investment trends. Her research covers software and infrastructure trends, offering insights that help leading vendors and infrastructure providers to develop strategy for future customer deployment scenarios.

Madsen has spent the past decade focusing on the datacenter and services market in Europe as an analyst and journalist. She most recently headed up CBRE's EMEA datacenter research, where she consulted with leading hyperscale and leased datacenter operators as well as the investment and real-estate developer community on adoption trends and strategy. She has had a seat at the table through the market's evolution as a result of cloud and regularly contributes to industry events as a speaker and moderator. Prior to CBRE she was a research director at S&P Global/451 Research, covering datacenter and hosted/managed services across EMEA. She also edited DatacenterDynamics' FOCUS website and magazine. Prior to this she ran multiple technology titles covering B2B in vertical sectors working for companies such as Progressive Media and CNet/ZDNet. She has a BA in journalism from Queensland University of Technology.

More about Penny Madsen

Message from the Sponsor

No matter where you are on your journey, Dell Technologies helps any organization change the way they can do business and become AI leaders.

Get in touch to learn how we can help you accelerate intelligent outcomes.

Visit Dell.com/AI to learn more

IDC Custom Solutions

This publication was produced by IDC Custom Solutions. As a premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets, IDC's Custom Solutions group helps clients plan, market, sell, and succeed in the global marketplace. We create actionable market intelligence and influential content marketing programs that yield measurable results.

IDC @idc @idc idc.com