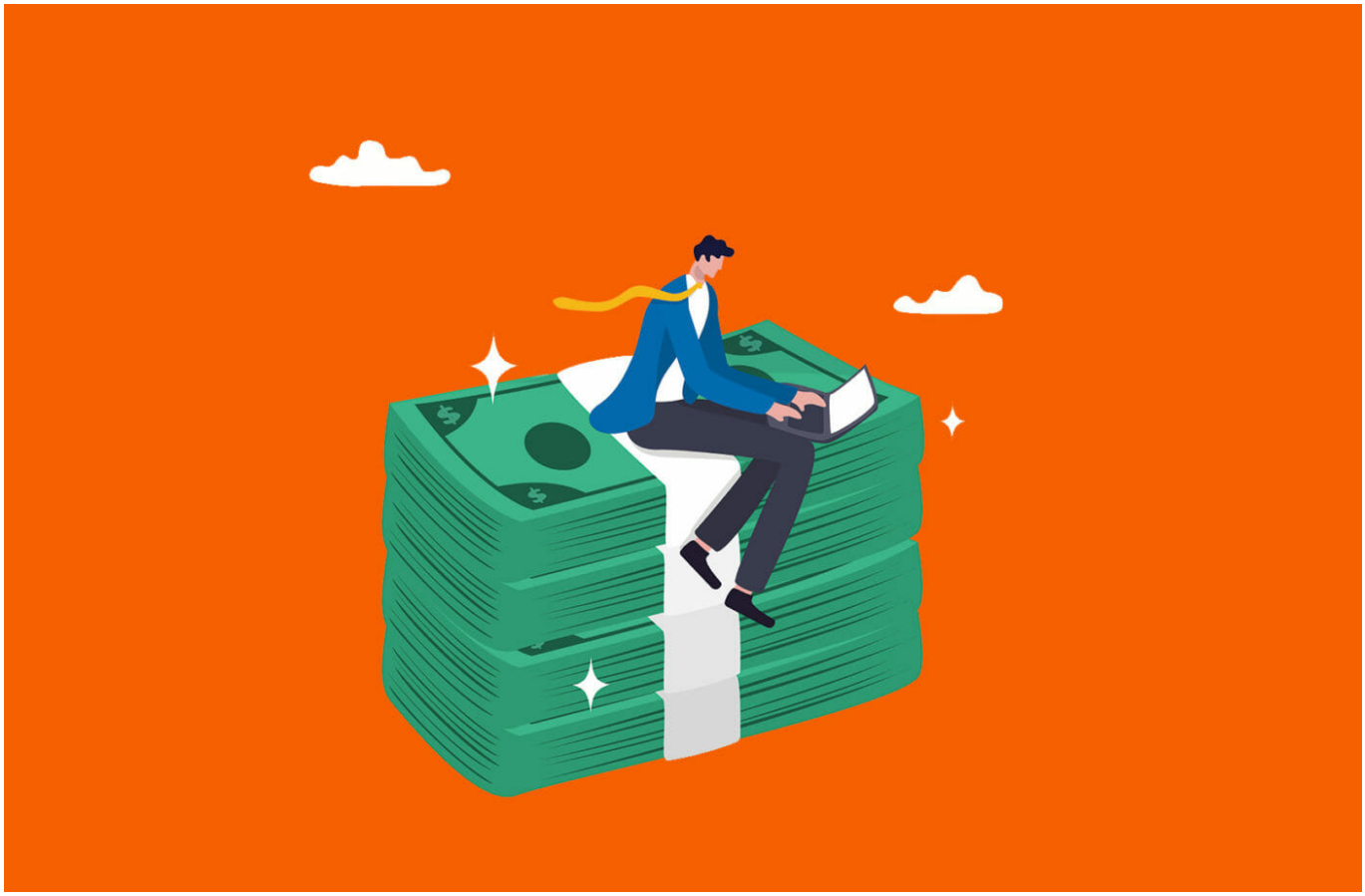
# How to Take Your AI into Production without Breaking the Bank



AI is like any other technology: The more you use it, the better and more helpful it becomes, but also, typically, the more expensive it becomes. Oftentimes, companies fail to consider what they'll do when their AI projects grow to the point where their total cost of ownership (TCO) starts to greatly exceed the value of the project.

AI projects can and usually do start small and from pretty much anywhere. Data scientists can work from a laptop, workstation, cloud resources, or from resources within powerful servers and storage in a data center.

The challenge comes when you scale your AI project. The initial resources for your AI computing may not be able to handle your bigger, more scaled-out AI efforts.

At some point, your production-level AI will need more horsepower and a more reliable infrastructure. But DIY solutions can be daunting and cloud AI solutions can be expensive (think ongoing rental costs of cloud compute, networking, and storage).

Here are some TCO considerations as you scale AI from exploratory pilot phases to production.

## The Allure, and Danger, of 'Big AI'

Everyone's familiar with the term "Big Data." Now there's "Big AI."

AI is growing very quickly and is becoming a top area of investment for enterprises globally.

Some stats that demonstrate the size and growth of AI:

- In the next five years, the global AI market value is forecasted to reach nearly $1.4 trillion.
- Investment in AI infrastructure is expected to double in the next four years.
- Within a few years, 88% of enterprises will have AI as either a major area of investment or largest area of investment.

When you survey the popular press to see what defines AI, you'll likely see three types of information:

- All the new and amazing capabilities that AI can bring to bear on a range of problems
- The various frameworks available to implement these solutions
- Cutting-edge hardware for those frameworks that continues to rapidly improve

The thing is, AI and machine learning (ML) are much more than this. If you're a

data architect or engineer responsible for industrial-scale AI, you know this very well.

And that's the danger of Big AI. It's kind of like gambling: Your eyes widen at the idea of what's possible and what you can win as any rational thought about what could go wrong or what you could lose goes out the window.

AI code is just a small fraction of real-world AI operations, as much of an AI system is dedicated to data collection, cleaning, labeling, verification, and management, in addition to infrastructure, which can be vast and complex.

Real-world AI and ML deployments can incur massive operational costs. The key is to consider and plan for the costs *before* you go into production.

## The Challenges of Taking AI from Pilot to Production

Your AI models may work great in the lab, but it's another story in production and at scale.

Large, production-scale data sets require scalable, high-performance compute and storage. Deployment flexibility requires apps built with portability in mind. Security, control, governance, and data ownership requirements require reliable, scalable performance on premises or in the cloud.

AI involves complex processes with multiple pipelines for data prep, model prototyping, training, and inference. Model development is also a non-linear process of exploration and experimentation. You're not building a single conventional app. It's a complex model based on a combination of analytical and machine learning algorithms. It's also an app, service, and collection of interfaces—all of which may change rapidly as needs and technologies evolve. And everything needs to be smoothly integrated if you actually want it to be accessible to end users who need its power.

Beware: Existing data center infrastructure and cloud resources may not meet the higher performance, scale, and/or availability requirements of production-level AI.

Lastly, accessing AI in production isn't a simple pass or fail of the code. It needs a data scientist involved to continually evaluate model performance—which can

degrade more rapidly than conventional software. So there's a lot of monitoring and retraining on a continual basis along the way. That's a big part of what makes it unlike conventional software, where a lot of enterprise IT and DevOps disciplines have evolved to industrialize software development.

The AI space requires a new, holistic approach where data science teams collaborate with DevOps and IT teams to get more models deployed in production.

## MLOps and the Reality of Scaling AI Projects

Let's say you have a great idea that provides the seed of an AI project, which turns into a model with decent accuracy, and you did this on your laptop or workstation. That's great! You got your inference working. You would save your example results, maybe write up a research paper, and you're done.

But if you wanted to now scale this project into production, you're probably going to have a bunch of ancillary concerns and supporting workloads running as well. Ideally, you planned for all these other activities from the beginning.

So, most of making enterprise AI successful is really about thinking ahead:

- How will you support all these important production activities?
- How will you avoid the pitfalls that can derail production?
- How will you maintain business value as patterns shift?

Taken together, the answers to the above questions combine into something known as machine learning operations, or "MLOps."

MLOps is planning for the infrastructure—the compute and storage components—that you're going to need as your AI project scales.

Of course, a key part of MLOps is data storage.

## AI Data Storage Requirements

The right storage platform can both simplify an AI deployment, enhance its value to your business, and keep your TCO lower than a traditional data center or cloud-only resources.

Data scientists' need for production data to adjust models and explore changing patterns and objectives argues strongly for a consolidated platform—a single, efficient, high-performance storage system that meets the needs of all project phases and allows development, training, and production to easily access dynamically evolving data.

Your data storage system should be capable of sustaining the intensive random access that concurrent training jobs require, even during occasional bursts of large sequential writes as jobs are checkpointed.

Your data storage system should also be capable of:

- Providing low latency for small random reads concurrently with high sequential write throughput
- Allowing data sharing by workstations and servers and rapid, non-disruptive capacity expansion
- Supporting parallel enumeration
- Delivering "24×365" operation throughout a project's life, self-healing when components fail and non-disruptive to expand and upgrade
- Protecting against human error (e.g., with data set snapshots) and fitting easily into data center ecosystems

Pure Storage® FlashBlade//S™ is a unified fast file and object system that sets new standards for performance, scalability, and simplicity in high-capacity file and

object data storage, making it perfect for AI. The all-flash blade-based system integrates hardware, software, and networking to offer higher storage density with lower power consumption and heat generation than other systems, along with versatile performance to support virtually any combination of file and object workloads.

*Learn more about leveraging FlashBlade//S for AI*.

*Schedule a meeting with Pure Storage and request an AI DGX Cloud vs. On-Prem TCO Analysis*