

# The AI/ML Startup's Guide to Overcoming Infrastructure Challenges

How to Optimize and Streamline Cloud Infrastructure  
With DevOps and FinOps



# Introduction

Artificial intelligence (AI) and machine learning (ML) startups are entering an era of unprecedented opportunity. In 2024, nearly one in three VC deals went to AI companies, with total funding [topping \\$100 billion](#) – an 80% leap from 2023. And AI/ML technology is advancing as rapidly as the business opportunities in the sector.

While generative AI (GenAI) has gone mainstream, a new frontier – agentic AI – is emerging. These systems, where multiple autonomous AI agents work together to perform multi-step processes and make complex decisions, are set to become foundational. [Gartner has projected that by 2028](#), 33% of enterprise software programs will include agentic AI, while it will autonomously make 15% of “day-to-day work decisions.”

But there is a catch. The more sophisticated your AI becomes, the more pressure it puts on your infrastructure. To get your startup’s offering to market, scale up, and be successful, you need to train, deploy, and iterate your models quickly, efficiently, and cost-effectively. That requires serious compute, flexibility, and cost control.

**All of that makes it increasingly important to streamline your infrastructure through effective DevOps and FinOps.**



**This eBook is your crash course in building the infrastructure maturity your AI/ML startup needs with the help of DevOps and FinOps principles. You will learn:**

- The challenges AI/ML startups face, both today and moving forward
- Best practices to streamline your infrastructure using DevOps and FinOps
- The key benefits of streamlined cloud infrastructure
- How to optimize your infrastructure as you scale

# Table of Contents

## Chapter 1: Key Factors for AI/ML Startups to Contend With

|                                                                |   |
|----------------------------------------------------------------|---|
| Rapid, data-intensive experimentation .....                    | 4 |
| Complex pipelines needed for machine learning .....            | 4 |
| Inconsistent, unpredictable, and expensive cloud usage .....   | 5 |
| Tool fragmentation and a lack of DevOps/FinOps expertise ..... | 5 |

## Chapter 2: 5 Tips for Streamlining AI/ML Infrastructure Using DevOps & FinOps

|                                                                   |   |
|-------------------------------------------------------------------|---|
| Tip 1: Embrace Infrastructure as Code (IaC) .....                 | 6 |
| Tip 2: Optimize Cloud Spend Without Sacrificing Performance ..... | 7 |
| Tip 3: Integrate Best DevOps Practices for ML Pipelines .....     | 7 |
| Tip 4: Leverage Cloud-Native AI/ML Tools .....                    | 8 |
| Tip 5: Consider Partnering for DevOps Expertise .....             | 8 |

|                  |   |
|------------------|---|
| Conclusion ..... | 9 |
|------------------|---|

# Key Factors for AI/ML Startups to Contend With

The potential for AI/ML startups is huge. But turning that potential into reality requires cloud infrastructure that meets your need for computing power while also being cost-effective, resilient, and adaptable.

Here are some of the most common factors that can complicate that goal for AI/ML startups:

## Rapid, data-intensive experimentation

Constant experimentation is at the heart of AI/ML. Models often go through dozens – or even hundreds – of iterations before reaching production. And startups may run hundreds of training jobs in a week, rapidly testing models and data variants to refine performance and iterate.

Agentic AI adds another layer of complexity, requiring continuous retraining and real-time feedback for multiple interacting agents.

To support this pace of innovation, AI/ML startups need extremely robust cloud infrastructure. It must handle huge amounts of data required for machine learning, adapt efficiently to increases in data processing, and run multiple parallel workloads without bottlenecks.

## Complex pipelines needed for machine learning

AI/ML development and deployment can be more complex than traditional software engineering. Deploying a machine learning model is an iterative, often unpredictable process, especially because of the need for continuous integration and continuous delivery (CI/CD) tailored to ML workflows.

Agentic AI raises the stakes even further. These systems rely on multi-agent architectures, where agents must communicate and coordinate in real time. If one agent is updated or fails, it can impact the behavior of others, making consistency and synchronization critical.

To manage this complexity, startups need robust deployment pipelines and rollback mechanisms that go beyond model versioning to mitigate the risk of unsuccessful updates.



### Inconsistent, unpredictable, and expensive cloud usage

By their nature, AI/ML workloads are unpredictable. Usage can spike suddenly – like when a training job encounters an unexpectedly large dataset, triggering a tenfold increase in GPU memory. These fluctuations make it hard to ensure that you have enough computing power without overspending.

For AI/ML startups with high memory and data-processing needs, rising cloud costs can be a challenge.

Gartner has projected:

**"Spending on public cloud services is expected to increase by 21.5% in 2025."**

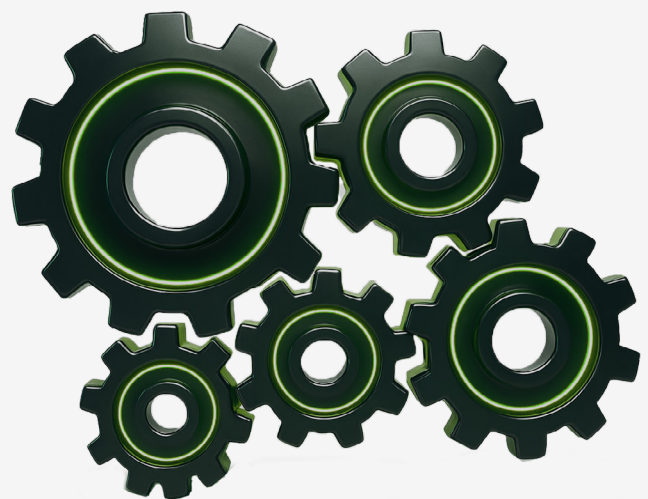
### Tool fragmentation and a lack of DevOps/FinOps expertise

It can be common for AI/ML startups' tech stacks to be more fragmented than in conventional software development. This patchwork approach can lead to broken workflows, slow debugging, and operational inefficiency.

Part of the challenge stems from relying on different tools at different stages, such as:

- PyTorch Lightning, Hugging Face Transformers, or TensorFlow/Keras for training.
- MLflow, Comet, or Weight & Biases for tracking experiments.
- Seldon Core, BentoML, or TorchServe for serving models.
- Prometheus, Arize AI, or Evidently AI for monitoring.

In addition, AI/ML startups are often founded by data scientists with expertise in modeling. While this provides them with elite knowledge in modeling to drive innovation, it often leaves them with limited experience in DevOps and FinOps. That can lead to missteps if your startup doesn't take steps to get access to sufficient DevOps and FinOps expertise.



# 5 Tips for Streamlining AI/ML Infrastructure Using DevOps and FinOps

The good news is that AI/ML startups' infrastructure challenges are solvable. The right DevOps and FinOps practices can make a powerful difference, helping you build infrastructure that scales with demand, remains cost-efficient, and supports iteration.

Aligning your cloud operations with technical and business goals helps you facilitate rapid experimentation and keep your products running smoothly, without compromising on efficiency and scalability.

**Here are five essential tips to help you overcome infrastructure challenges using DevOps and FinOps:**

## Tip 1: Embrace Infrastructure as Code (IaC)

Infrastructure as code (IaC) lets you define and manage cloud infrastructure using configuration files to provision resources automatically, rather than setting up servers, databases, and networks manually.

This script-based approach delivers the speed, consistency, and repeatability AI/ML startups crave by letting you reuse your specifications across various environments. For example, you can define the architecture for a cluster of agents once in Terraform and replicate it across staging, testing, and production environments.

### IaC also enables:

- Version control for infrastructure, making it easier to trace results back to specific configurations.
- Automation that reduces errors like misconfigured virtual machines.
- The earlier you adopt IaC, the better. Starting from day one (or as close to it as possible) helps gain agility and prevent chaos as you scale – especially when dealing with highly complex and dynamic environments like agentic AI architecture.

### Tip 2: Optimize Cloud Spend Without Sacrificing Performance

To make models and products run smoothly without breaking the bank, you need just the right amount of cloud computing power. DevOps and FinOps help you strike the balance between delivering performance and avoiding waste.

#### Three essential strategies to do that are:

- **Auto-scaling** – automatically increasing (or decreasing) your cloud computing power in real time based on demand. This allows you to set rules for scaling infrastructure depending on spikes or lulls.
- **Right-sizing** – analyzing historical trends in your data usage to align resource allocation with actual needs.
- **Using spot instances** – discounted virtual machines offered at off-peak times. This is ideal for non-critical or batch workloads.

### Tip 3: Integrate Best DevOps Practices for ML Pipelines

For any effort to streamline an AI/ML startup's infrastructure, it is essential to follow best DevOps and MLOps practices. And with agentic AI raising the bar for computing power even further, the principles of DevOps are now becoming mission-critical in this field.

Specifically, using DevOps practices to overcome AI/ML infrastructure challenges involves:

- Emphasizing CI/CD for rapid testing, deployment, and safe rollbacks if necessary.
- Automation for monitoring, so that models and agents perform reliably and securely across environments.



### Tip 4: Leverage Cloud-Native AI/ML Tools

Given AI/ML startups' heavy reliance on cloud infrastructure, using cloud-native tools is key to making that infrastructure work harder, smarter, and as efficiently as possible. These tools allow your team to focus less on managing logistics and more on building models and products that matter.

Cloud-native AI/ML tools offer several advantages:

- **Smooth integration and auto-scaling.**

Cloud-native tools adapt in real time to your needs, helping you scale automatically and pay only for what you use – without the overhead of manual scaling.

- **Simplified infrastructure management and maintenance.** Relieve your team of complex tasks like cluster provisioning, managing GPU drivers, and maintaining environments.

- **Built-in security.** With infrastructure managed at scale by your cloud provider, you gain enterprise-grade security and stability without additional complexity.

### Tip 5: Consider Partnering for DevOps Expertise

Many AI/ML startups face a shortage of DevOps and FinOps expertise. It's understandable that startups in this space tend to prioritize the hiring of data science specialists and engineers at the outset as they develop core products, but this can create knowledge gaps around infrastructure management.

As your startup scales, those gaps can cause bottlenecks, inefficiencies, or even outages.

Supplementing your in-house talent with on-tap DevOps expertise offers a quick, flexible, and robust solution.

DevOps as a Service is a particularly beneficial model for AI/ML startups, providing immediate access to prebuilt tools and a wealth of DevOps knowledge so you can:

- Accelerate your time to market
- Reduce cloud costs
- Avoid infrastructure pitfalls

Automat-it's DevOps-as-a-Service clients have enjoyed up to [70% reductions in cloud cost spend](#), improved time to market [by 50%](#) and dramatically reduced [data storage costs](#).

Perhaps most importantly, outsourcing DevOps to trusted professionals allows your startup team to focus more on gathering data, refining models, and optimizing product offerings.

# Conclusion

After massive growth in recent years, the AI market is projected to soar from \$371.7 billion in 2025 to \$2.4 trillion in 2032 – [a combined annual growth rate of 30.6%](#). To seize this opportunity, AI/ML startups must streamline their cloud infrastructure.

The key is leveraging DevOps and FinOps to overcome common infrastructure challenges. Doing so means waving goodbye to being bogged down by operational hurdles – and welcoming in a more efficient, focused era of innovation. That involves:



Infrastructure as code (IaC) for reproducible, scalable environments.



Cost optimization to ensure that your company has adequate cloud computing power without overprovisioning and overpaying.



Automated, streamlined CI/CD pipelines to avoid problems as your company deploys and improves its products.



Cloud-native tools to optimize those products' performance and maximize their efficiency



Partnerships to meet your company's need for DevOps and FinOps expertise quickly, flexibly, efficiently, and scalably.

By combining these five strategies, you can enjoy faster experimentation and model deployment, greater system stability, lower cloud costs, and fewer technical headaches. This approach can accelerate innovation and help you scale your cloud infrastructure as you grow. Ultimately, that means a clearer path to game-changing AI/ML products and sustainable success.

With optimized cloud infrastructure, your AI/ML startup can move faster, spend smarter, and scale bigger. Automat-it can streamline that infrastructure for you, letting you enjoy the speed, flexibility, and efficiency of DevOps as a Service. Get in touch to find out how.

# Who we are

Automat-it is an all-in AWS Premier Partner providing professional services with deep domain expertise to achieve your KPIs from business growth to availability and cost control.

Our DevOps DNA with deep AWS expertise ensures that we bring practical solutions that foresee opportunities and challenges to save valuable time to market and resources throughout your growth journey.

## Company information:

**400+**

AWS certifications

**600+**

Customers

**120+**

Engineers

**9.6**

CSAT for 2025

[Contact us](#)