RESEARCH REPORT

Data on Kubernetes 2025

Beyond Adoption: The Age of Operational Excellence

NOVEMBER 2025

DoK Community

Executive Summary

After five years of surveying the Data on Kubernetes (DoK) community, the data tells a definitive story: DoK has won. The architectural battle is over—running data workloads on Kubernetes is now standard practice. What's top of mind for organizations today is operational excellence: how to optimize DoK deployments, manage costs, and prepare for the explosive growth of AI/ML workloads.

Our 2025 survey of 182 technology professionals reveals that the community has crossed the chasm from adoption to optimization. Organizations aren't asking "Should we run data on Kubernetes?" but rather "How do we excel at it?" This fundamental shift is reflected in every metric: from production deployment patterns to strategic priorities to the challenges organizations face.

The data reveals four critical trends shaping the future of DoK:

1. **Cost Optimization Becomes the Top Priority:** Optimizing costs has emerged as the #1 priority for 2025, surpassing AI/ML improvements, security enhancements, and scaling initiatives. This is particularly acute for organizations running AI/ML workloads, where storage costs (50%) have become the primary concern—reflecting the enormous data requirements of training datasets, model checkpoints, and inference results for large-scale AI deployments.

2. **The AI/ML Revolution Accelerates:** While databases maintain their #1 position (66%), AI/ML workloads have surged to 44% adoption. More striking is what's happening beneath the surface: vector databases are seen as critical infrastructure by 77% of respondents—the strongest signal in our survey. The RAG (Retrieval-Augmented Generation) revolution is here, and it's reshaping data infrastructure requirements.

3. **The Edge + Real-time Architectural Shift:** Organizations are making a decisive move toward distributed architectures: 61% view edge computing as essential, while 64% say real-time data processing is critical for their AI strategy. This represents a fundamental shift from centralized, batch-oriented workloads to distributed, real-time systems.

4. **Performance Gaps Reveal Optimization Opportunities:** Despite widespread adoption, performance bottlenecks persist. Storage I/O performance is cited as the primary concern, followed closely by model/data loading times. These gaps represent both challenges and opportunities for the ecosystem to deliver better tooling, practices, and infrastructure.

The maturity of DoK is evident in the production statistics: Nearly half of organizations run 50% or more of their DoK workloads in production, with the most advanced running 75%+ in production. Organizations attribute significant business value to DoK, with 62% linking 11% or more of their revenue to these deployments.

However, maturity brings new challenges. The top operational concerns are no longer about basic adoption but about optimization: performance optimization (46%), security and compliance (42%), and talent/skills gaps (40%). The skills gap is particularly acute—organizations need practitioners who understand both Kubernetes operations AND data workload optimization.

Looking ahead, organizations are prioritizing cost optimization (25%) and AI/ML improvements (24%) for 2025. The majority plan to focus on "selective optimization of current deployments" rather than aggressive expansion—a clear signal that the focus has shifted from growth to efficiency.

This report provides technology leaders with insights into how the most advanced organizations are optimizing DoK, overcoming operational hurdles, and positioning themselves for the AI-driven future. The message is clear: DoK is now the foundation, and operational excellence is the differentiator.

Key Findings

DoK Has Reached Production Maturity

- **Nearly half organizations run 50% or more** of their DoK workloads in production
- **62% attribute 11%+ of revenue** to running data on Kubernetes
- **More than half report productivity gains of 50% or more,** with 19% reporting gains of 2x or more

Databases Remain the Foundation, AI/ML is the Future

- **Databases maintain #1 position** for the fourth consecutive year
- **AI/ML workloads surge to 44% adoption,** up significantly from previous years
- **Vector databases emerge as critical infrastructure** with 77% seeing them as essential—the strongest signal in the survey

Cost Optimization Becomes the Top Priority

- **Optimizing costs is the #1 priority for 2025,** which is particularly acute for organizations running AI/ML workloads where storage costs dominate
- **Organizations are implementing multiple cost strategies:** GPU utilization optimization (50%), storage tier optimization (46%), cross-zone data transfer reduction (40%)
- **For those running AI/ML workloads,** they implement auto-scaling (58%), resource tagging (44%), storage tier optimization (46%)

The Edge + Real-time Architectural Shift

- **61% view edge computing as essential** for their future data strategy
- **64% say real-time data processing is critical** for their AI strategy
- This represents a fundamental move away from centralized, batch-oriented architectures

Performance Gaps Reveal Opportunities

- **Storage I/O performance is the #1 performance bottleneck (24%),** followed closely by model/data loading times (23%), indicating that data access patterns are the primary constraint for DoK workloads
- **Organizations implement numerous storage strategies:** Object storage integration (43%), local SSDs for performance (43%), caching layers (42%), block storage (42%)
- **Organizations running AI/ML workloads are further deploying multiple storage acceleration techniques:** in-memory caching (43%), local SSD caching (36%), model streaming (35%)

Training Dominates Over Inference

- **58% focus primarily on training workloads** vs 42% on inference
- This contradicts conventional wisdom that inference would dominate—organizations are still building and fine-tuning models

Skills Gap Persists Despite Maturity

- **40% cite talent/skills gaps** as a top operational challenge
- The gap is acute: organizations need practitioners who understand both Kubernetes AND data workload optimization

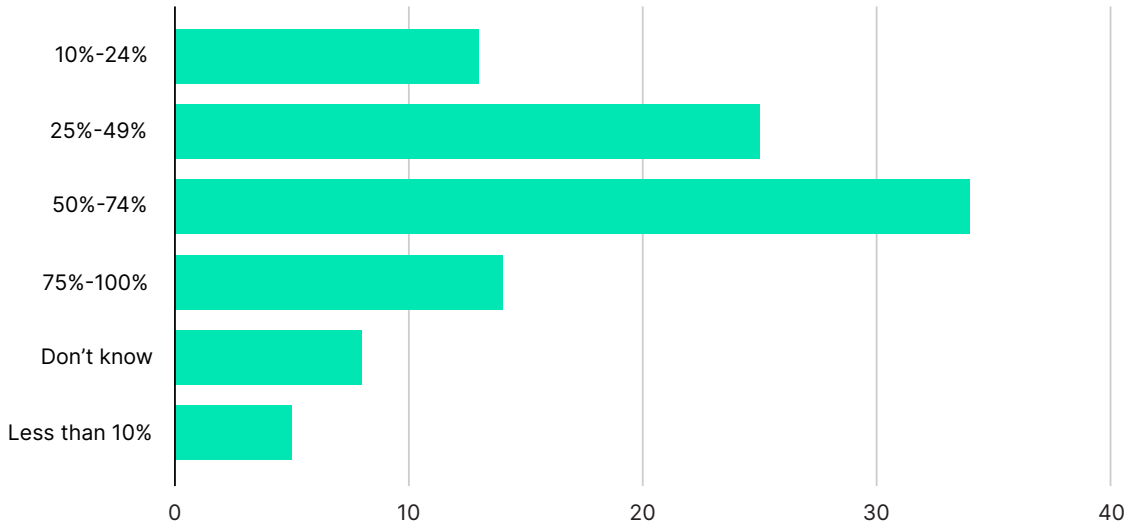
The Maturation of DoK: From Adoption to Excellence

Production Deployment Patterns

The journey from 2021 to 2025 shows clear maturation. In 2021, 70% were running stateful workloads. Today, the story is about scale and sophistication:

Production DoK Workloads

What percentage of your data on Kubernetes workloads are running in production?



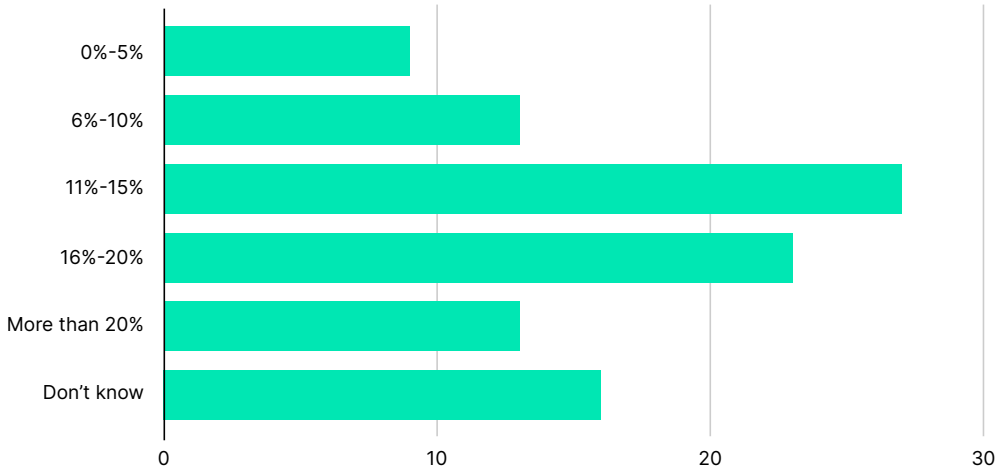
Nearly half of organizations have reached the critical threshold of running 50% or more of their data workloads on Kubernetes. This isn't experimental—it's mission-critical infrastructure.

Business Impact and Productivity

Organizations continue to see significant business value from DoK deployments:

Revenue Attribution to DoK

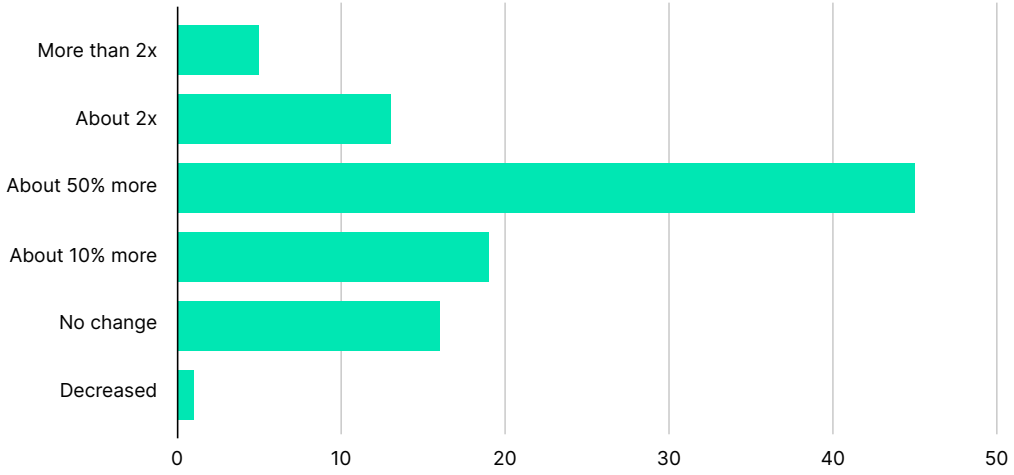
What percentage of your organization's revenue would you attribute to running data on Kubernetes?



Combined, 62% of organizations attribute 11% or more of their revenue to DoK—demonstrating clear business impact beyond engineering efficiency.

Productivity Improvements

How much has running DoK improved your organization's productivity?



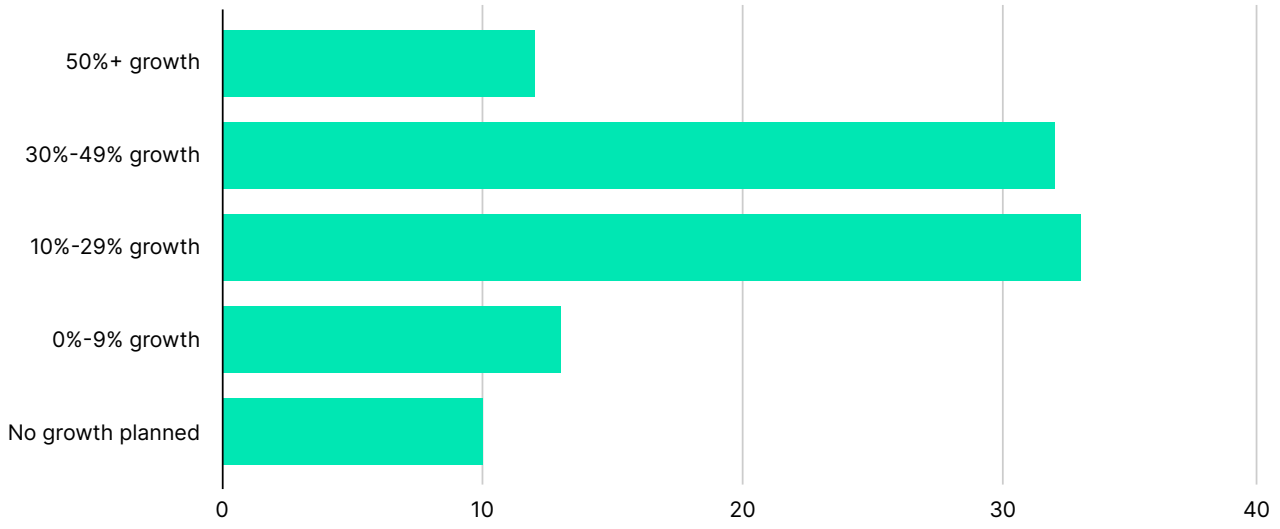
The productivity story remains strong, with the majority seeing significant gains. The 19% achieving 2x or greater productivity represents organizations that have mastered DoK operations.

Growth Expectations

Despite the maturity, growth continues:

Expected DoK Growth in 2025

By how much do you expect your DoK footprint to grow in 2025?



Three-fourths of the respondents expect at least 10% growth, signaling that even mature deployments continue to expand their DoK footprint.

Workload Analysis: What's Running on Kubernetes

The Top Five

1. **Databases:** 66% For the fourth consecutive year, databases remain the #1 workload on Kubernetes. This consistency demonstrates the platform's reliability for mission-critical data services and validates that the concerns about running stateful workloads in Kubernetes have been thoroughly addressed.

2. **Analytics:** 57% Analytics has emerged as the #2 workload, reflecting the growing sophistication of data platforms. Organizations are running complex analytical workloads—from data warehouses to data lakes—directly on Kubernetes.

3. **AI/ML:** 44% AI/ML's rise to the #3 position represents the most significant shift in the DoK landscape. This workload type is driving infrastructure innovation, cost concerns, and architectural decisions across the ecosystem.

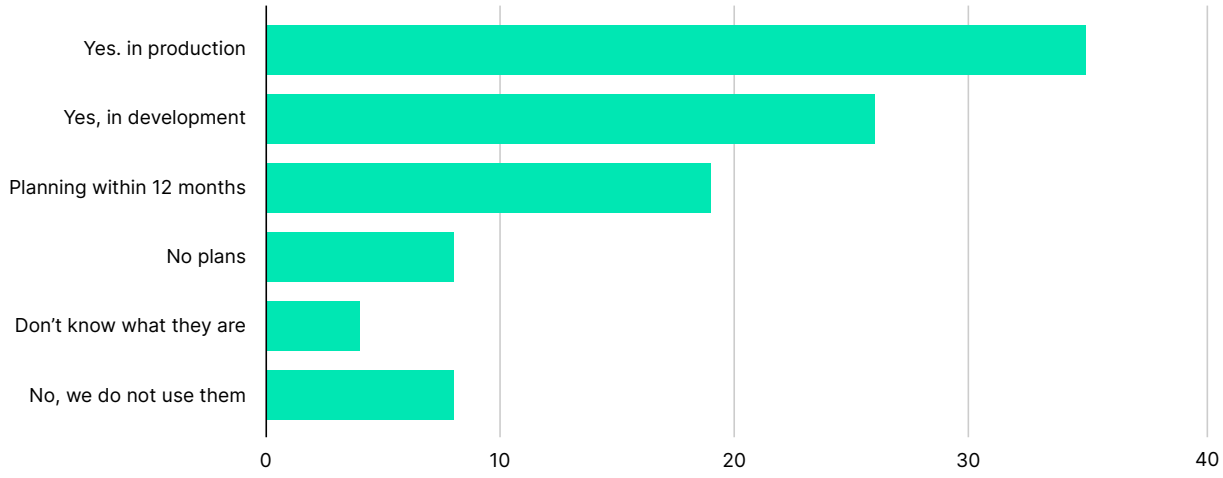
4. **Streaming/Messaging:** 41% Real-time data streaming has become a core capability, supporting use cases from operational analytics to AI/ML feature pipelines.

5. **Real-time Processing:** 40% Closely related to streaming, real-time processing reflects the industry shift from batch to stream-oriented architectures.

The Emerging Category: Vector Databases

Vector Database Adoption

Do you use vector databases as part of your DoK workloads?



Combined, 80% are either using or planning to use vector databases—a remarkable adoption curve for an emerging technology. This reflects the rapid adoption of RAG architectures for AI applications.

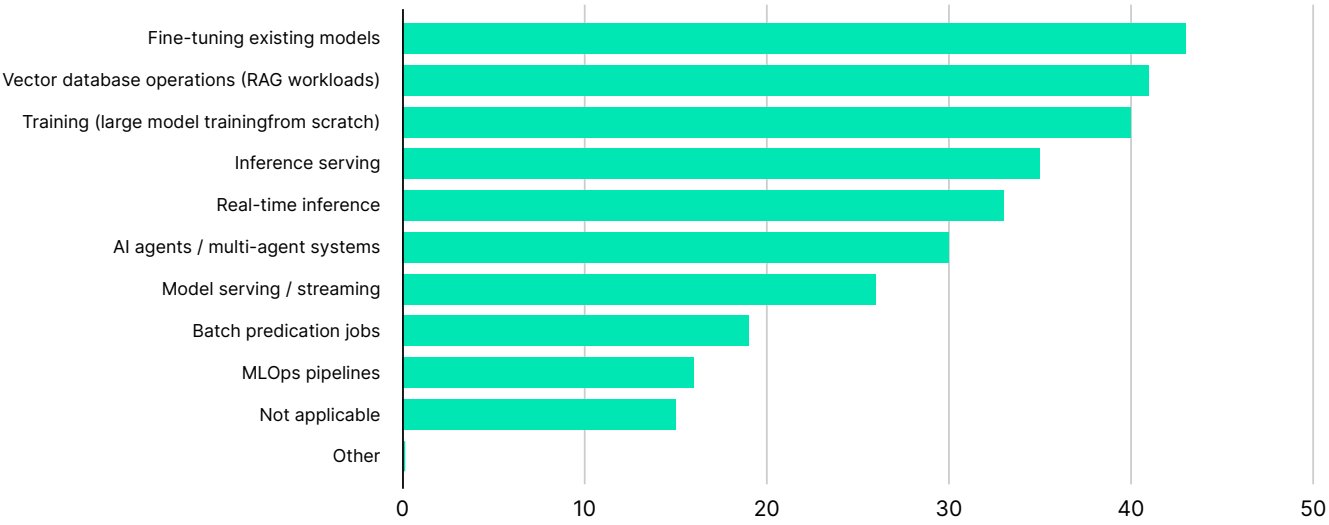
The AI/ML Revolution on Kubernetes

AI/ML Workload Patterns

Organizations are running diverse AI/ML workloads on Kubernetes:

Top AI/ML Workloads

Which AI/ML workloads do you run on Kubernetes? (Select all that apply)



The dominance of fine-tuning and training over inference is noteworthy. Despite industry predictions that inference would dominate, organizations are still heavily focused on building and customizing models. This likely reflects:

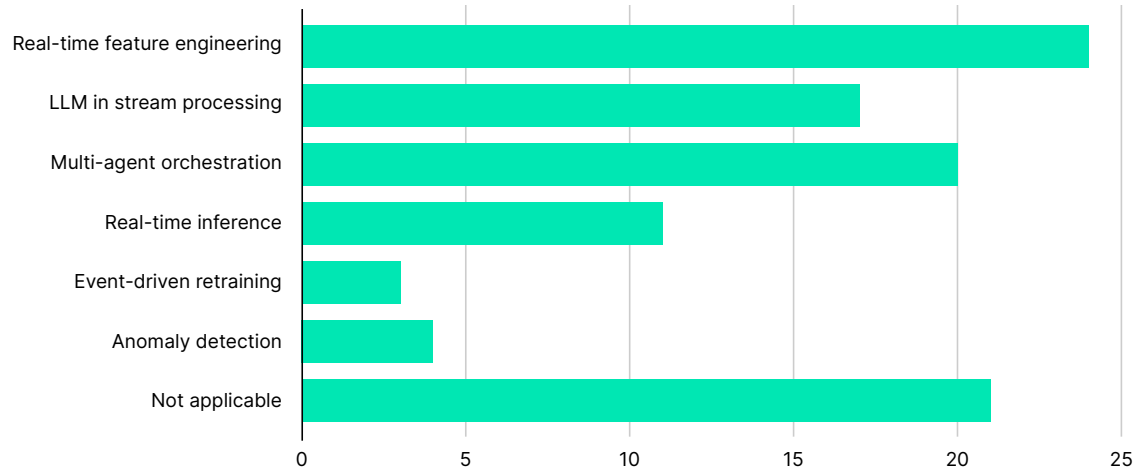
1. The rapid pace of model evolution requiring continuous retraining.
2. The need for domain-specific fine-tuning.
3. Organizations building proprietary models rather than solely consuming pre-trained ones.

The RAG Architecture Revolution

The surge in vector database adoption (77% viewing them as critical infrastructure) represents the fastest adoption of any infrastructure component we've tracked. This is driven by:

RAG Use Cases

If you use streaming and AI/ML: How do you integrate them? (Select all that apply)

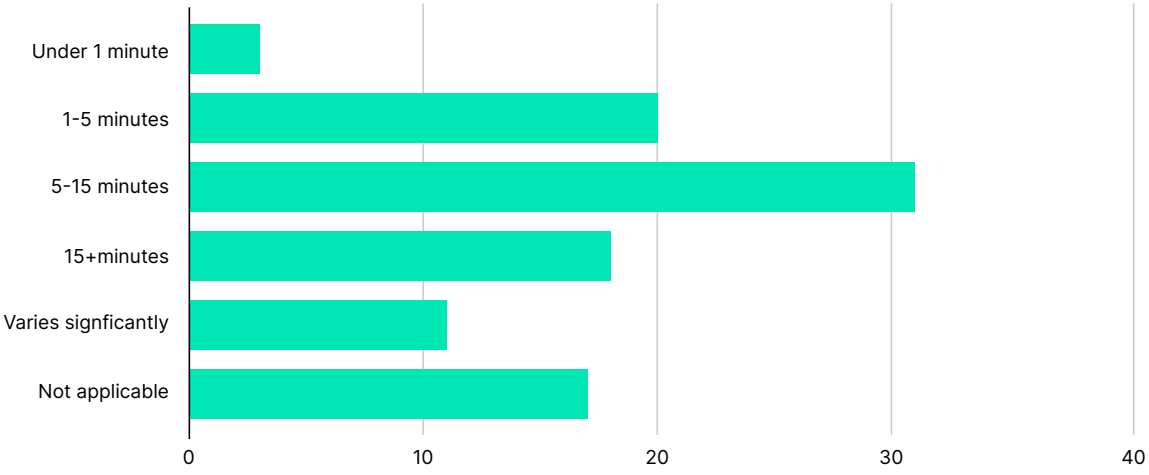


RAG architectures have become the dominant pattern for production AI applications, enabling organizations to augment LLMs with domain-specific knowledge without costly retraining.

AI/ML Performance Challenges

Model Loading Times

If you use AI/ML: What's your typical model loading time to ready-to-serve?



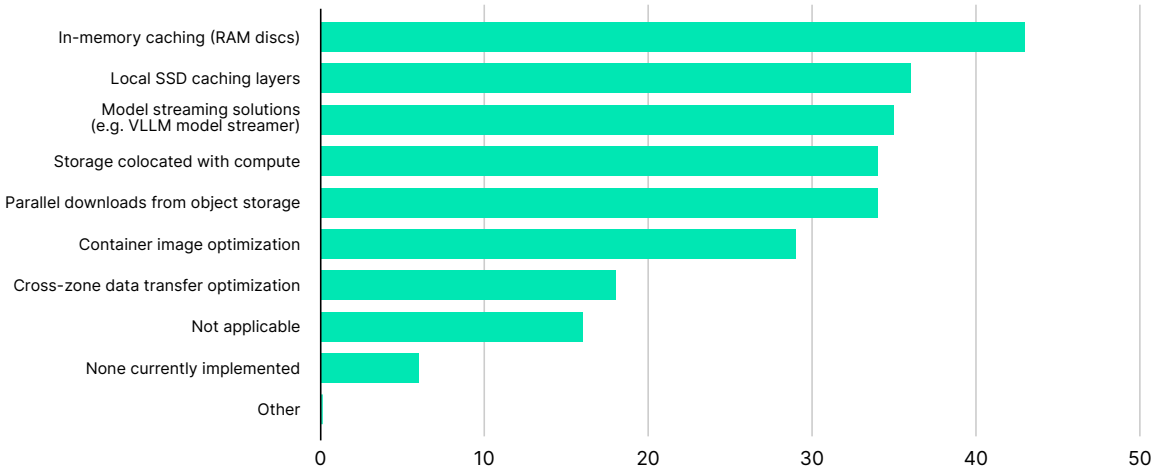
Only 3% achieve sub-1-minute model loading—a critical performance gap. With GPU costs often exceeding $1/hour per GPU, even 5-minute loading times represent significant waste. This is driving adoption of storage acceleration techniques.

Storage Acceleration Strategies

Organizations are deploying multiple strategies simultaneously:

Storage Acceleration Techniques

What storage acceleration techniques do you use for AI/ML workloads? (Select all that apply)

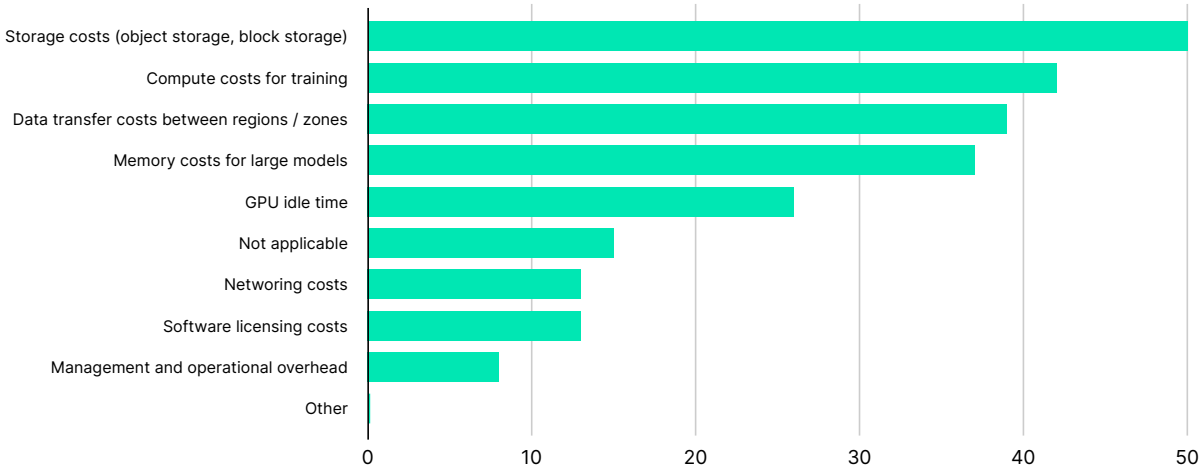


The diversity of techniques in use suggests the ecosystem is still experimenting to find optimal approaches for different workload types.

AI/ML Cost Challenges

Top Cost Concerns

If you use AI/ML: What is your biggest cost concern with AI/ML workloads on Kubernetes? (Select top THREE)



The emergence of storage costs as the #1 concern represents a significant shift. As models grow larger and training datasets expand, the cost of storing and accessing data has become a primary cost driver—sometimes exceeding GPU costs themselves.

GPU Utilization Optimization Priority

- Extremely important (top priority/KPI): 40%
- Very important (actively monitor/optimize): 28%

Combined, 68% rate GPU optimization as extremely or very important, reflecting the high cost of GPU infrastructure and the business imperative to maximize utilization.

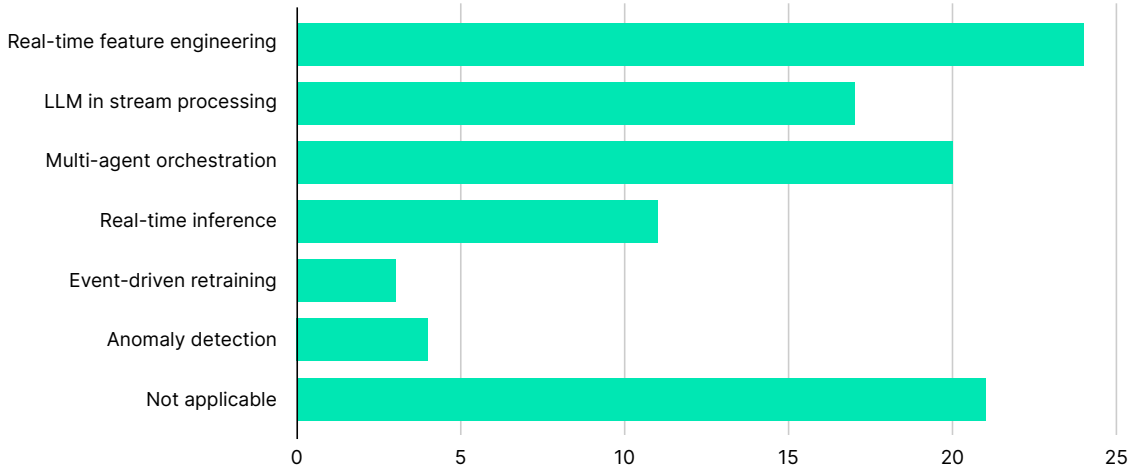
Streaming Data Integration with AI/ML

Streaming + AI/ML Integration

- Yes, extensively: 35%
- Yes, specific use cases: 27%

62% are integrating streaming data with AI/ML workloads, enabling:

If you use streaming and AI/ML: How do you integrate them? (Select all that apply)



This integration represents the convergence of two major trends: real-time data processing and AI/ML, creating systems that can make intelligent decisions in real-time.

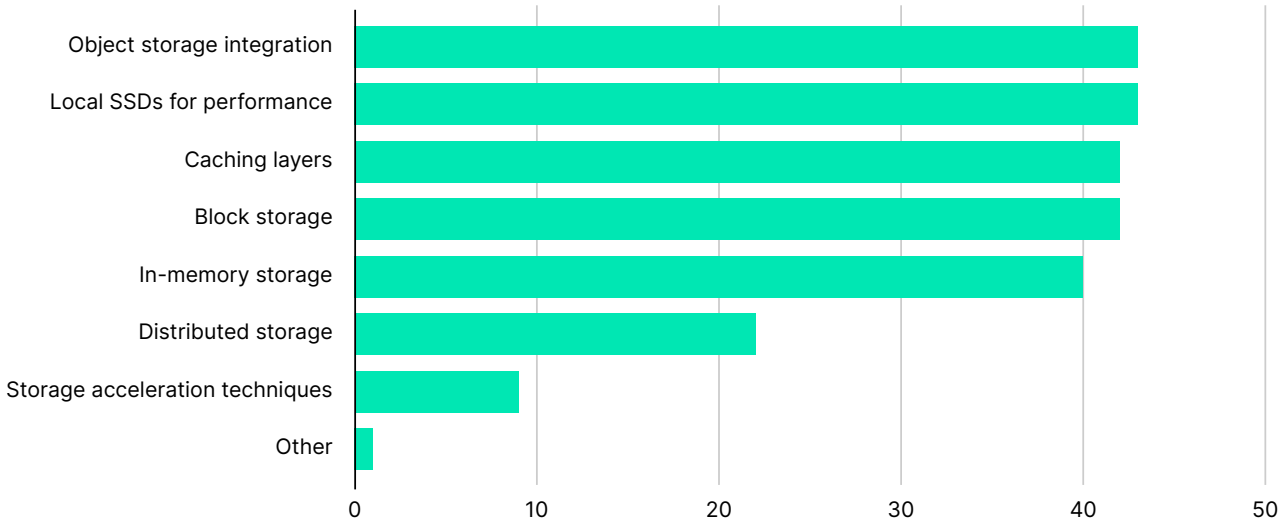
Storage and Performance Optimization

Storage Strategies in Production

Organizations are using diverse storage strategies, often multiple simultaneously:

Primary Storage Strategies

What storage strategies do you use for DoK? (Select all that apply)

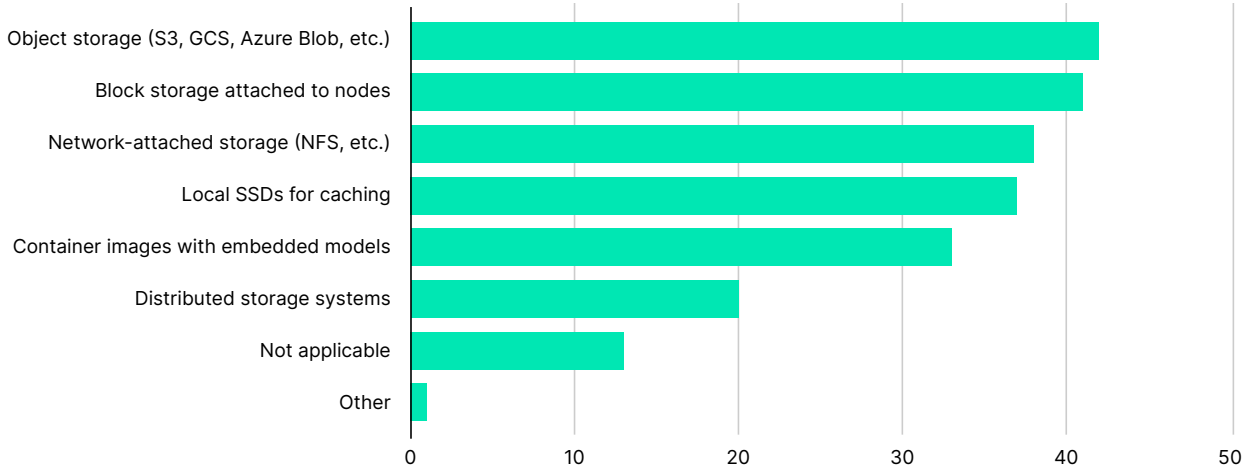


The balance between object storage (cost-effective, scalable) and local SSDs (high-performance) reflects the dual priorities of cost management and performance optimization.

Model Storage and Loading

How Organizations Handle Large Models

How do you primarily handle large model storage and loading? (Select all that apply)

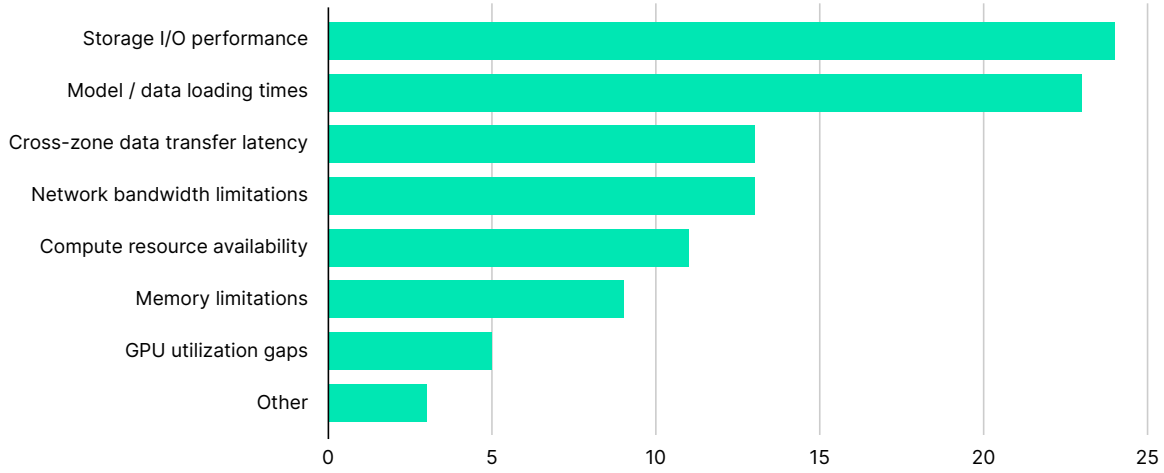


The prevalence of object storage (42%) explains the performance challenges—object storage, while cost-effective and scalable, has higher latency than local storage options.

Performance Bottlenecks

Biggest Performance Bottlenecks

What is your biggest performance bottleneck with data workloads on Kubernetes?



Storage and data movement dominate the bottleneck list, validating the focus on storage acceleration techniques.

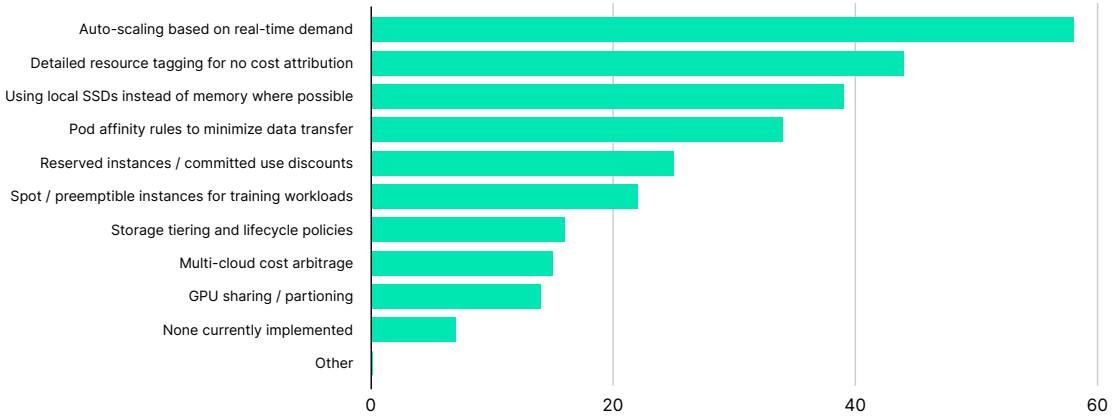
Cost Management: The New Frontier

Optimizing costs emerged as the #1 DoK priority for organizations.

Cost Optimization Strategies in Practice

Most Implemented Strategies

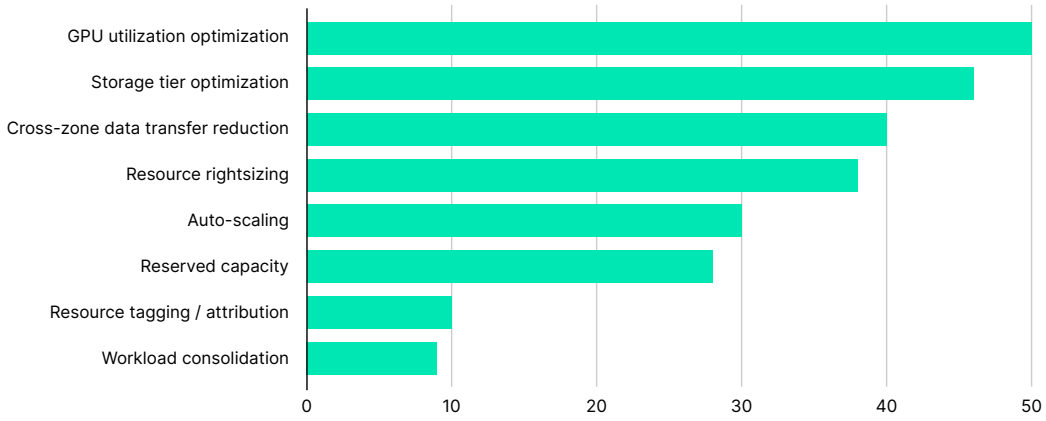
Which cost optimization strategies do you currently implement? (Select all that apply)



Auto-scaling (58%) is the most widely implemented strategy, reflecting its effectiveness at matching resources to actual demand.

Top 3 Cost Optimization Priorities

How do you primarily optimize costs for DoK? (Select top 3)



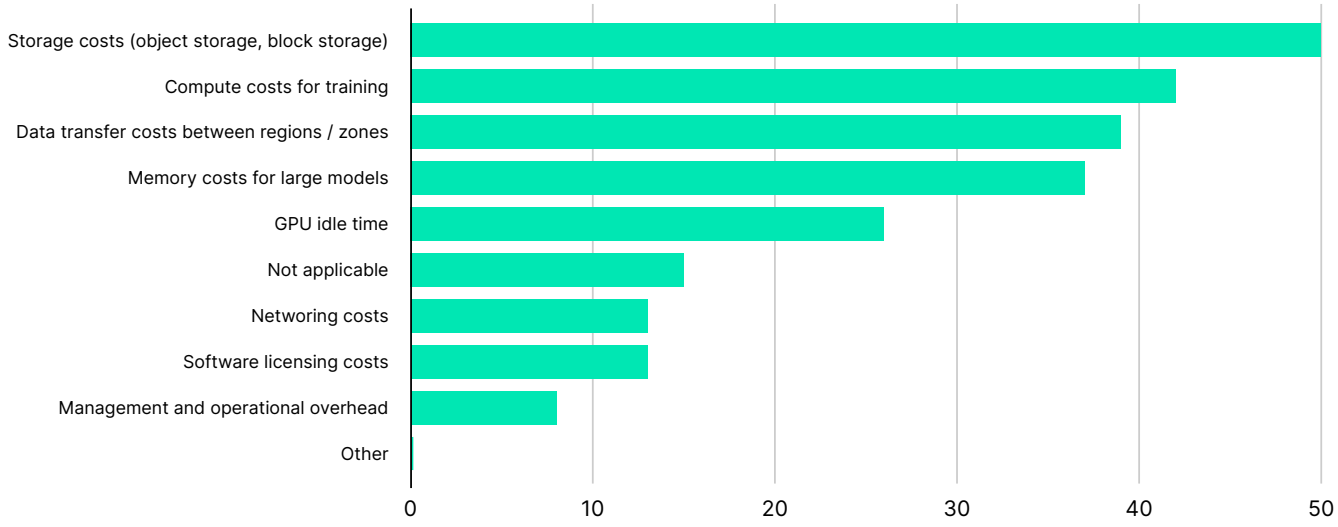
The focus on GPU utilization, storage, and data transfer reflects the three major cost drivers for DoK workloads.

AI/ML Top Cost Concerns

The cost landscape has shifted dramatically:

Primary Cost Concerns (AI/ML Workloads)

If you use AI/ML: What is your biggest cost concern with AI/ML workloads on Kubernetes? (Select top THREE)



Storage costs have emerged as the dominant concern, reflecting:

- Growing model sizes (70B+ parameter models)
- Large training datasets
- Need for multiple model versions
- Replication across regions/zones

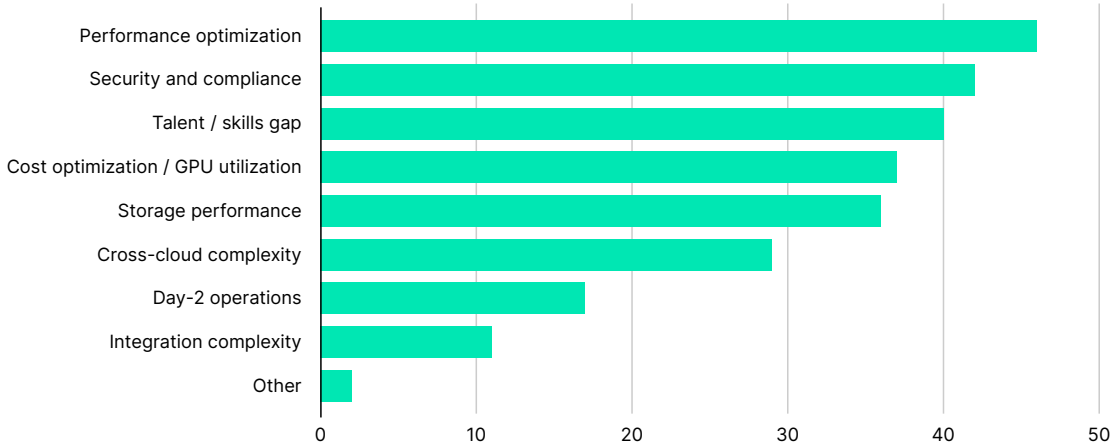
Operational Challenges and Governance

Top Operational Challenges

The nature of challenges has evolved from adoption to optimization:

Top 3 Operational Challenges

What are your TOP 3 operational challenges with DoK today? (Select up to THREE)



Performance optimization has emerged as the #1 challenge, displacing earlier concerns about basic readiness or vendor solutions. This reflects organizational maturity—the question is no longer "can we run it?" but "how do we run it optimally?"

The Persistent Skills Gap

Talent/Skills Gaps: 40%

Despite years of DoK growth, the skills gap remains a top 3 challenge. Organizations need practitioners who understand:

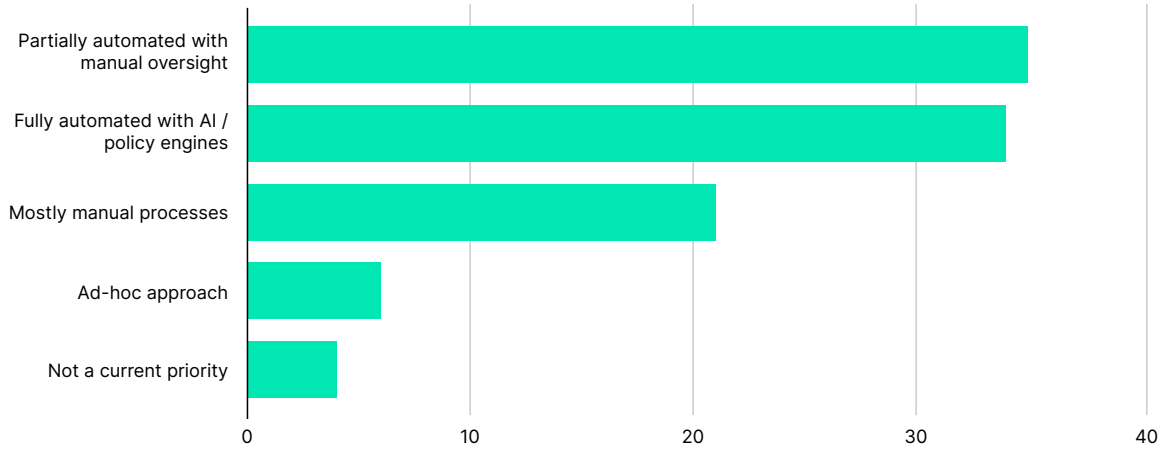
- Kubernetes operations
- Data workload optimization
- Storage performance tuning
- Cost management
- AI/ML infrastructure requirements

This represents a significant opportunity for training, certification, and tooling that simplifies operations.

Data Governance Approaches

Data Governance for DoK

How do you view data governance for DoK workloads?

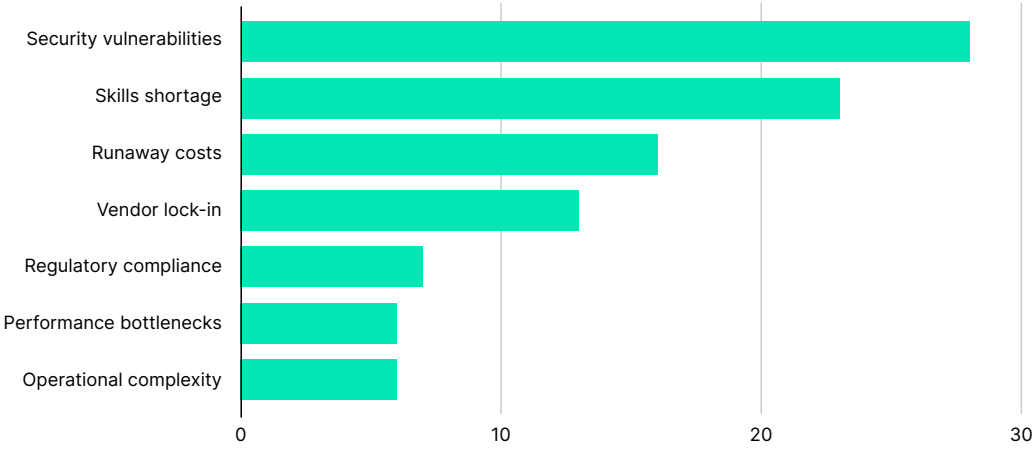


69% have implemented at least partial automation for data governance, reflecting the complexity of managing data across distributed Kubernetes environments.

Biggest Concerns for the Next Year

Top Concerns

What's your biggest concern about DoK in the next year?



Security has emerged as the #1 concern, likely driven by:

- High-profile Kubernetes security incidents
- Complexity of securing distributed data workloads
- Regulatory compliance requirements
- AI/ML data sensitivity

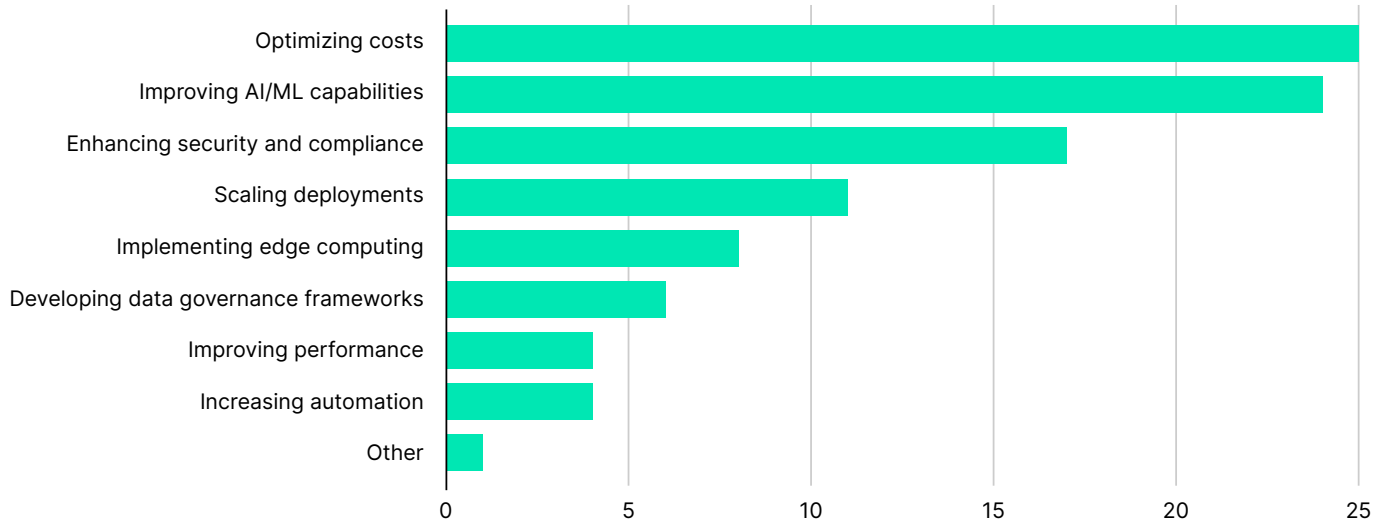
Strategic Outlook: The Future of DoK

2025 Priorities

Organizations are focused on optimization and capability enhancement:

#1 Priority for 2025

What's your #1 DoK priority for 2025?

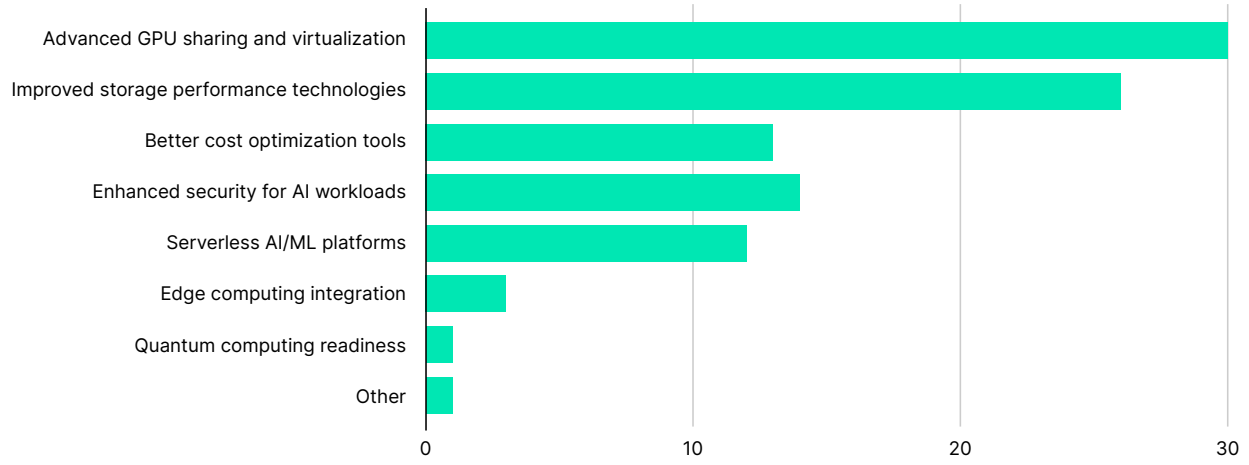


The combined focus on cost optimization and AI/ML improvements (49%) reflects the dual imperatives of efficiency and innovation.

Emerging Technology Trends

Technologies That Will Most Impact DoK

Which emerging technology trend will most impact your data on Kubernetes deployments in the next 2 years?

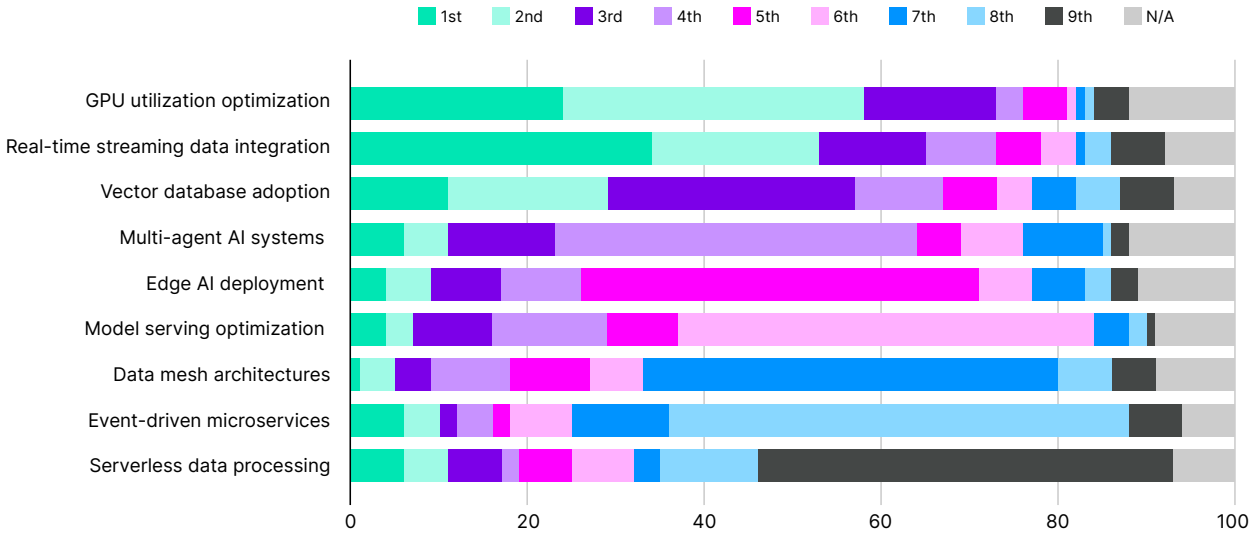


GPU optimization and storage performance dominate, validating the survey's focus on these areas.

Architectural Patterns for 2025

Most Important Patterns (Top Rankings)

Which of the following architectural patterns do you see as most important for your data on Kubernetes strategy in 2025? Please rank in order of importance.

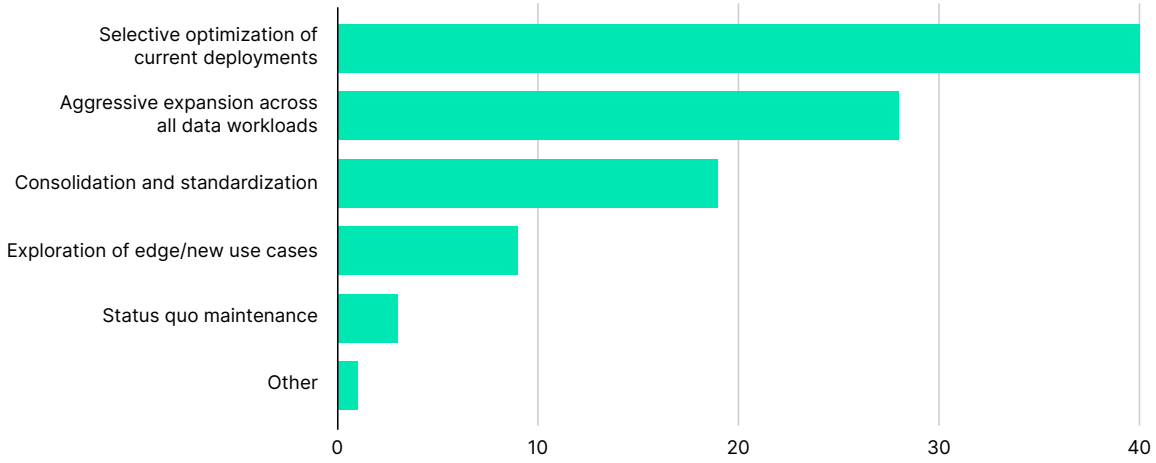


The diversity of #1 rankings suggests organizations are pursuing different strategies based on their specific use cases, but all are focused on AI/ML optimization.

DoK Evolution Path

How Organizations See DoK Evolving

How do you see DoK evolving in your organization over the next 2 years?

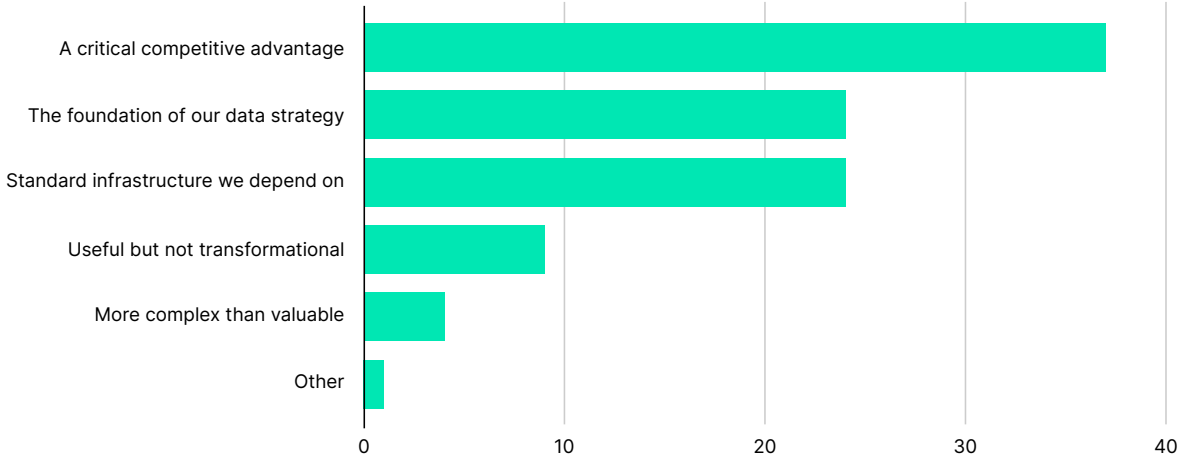


The majority (40%) planning "selective optimization" confirms that the focus has shifted from growth to efficiency. Organizations are saying: "We've adopted DoK, now let's make it excellent."

Strategic Positioning

86% view DoK as either foundational or a competitive advantage — strong validation that DoK has moved from experimental to strategic

Complete this statement: "Data on Kubernetes has become..."



Strategic Insights: Four Critical Trends

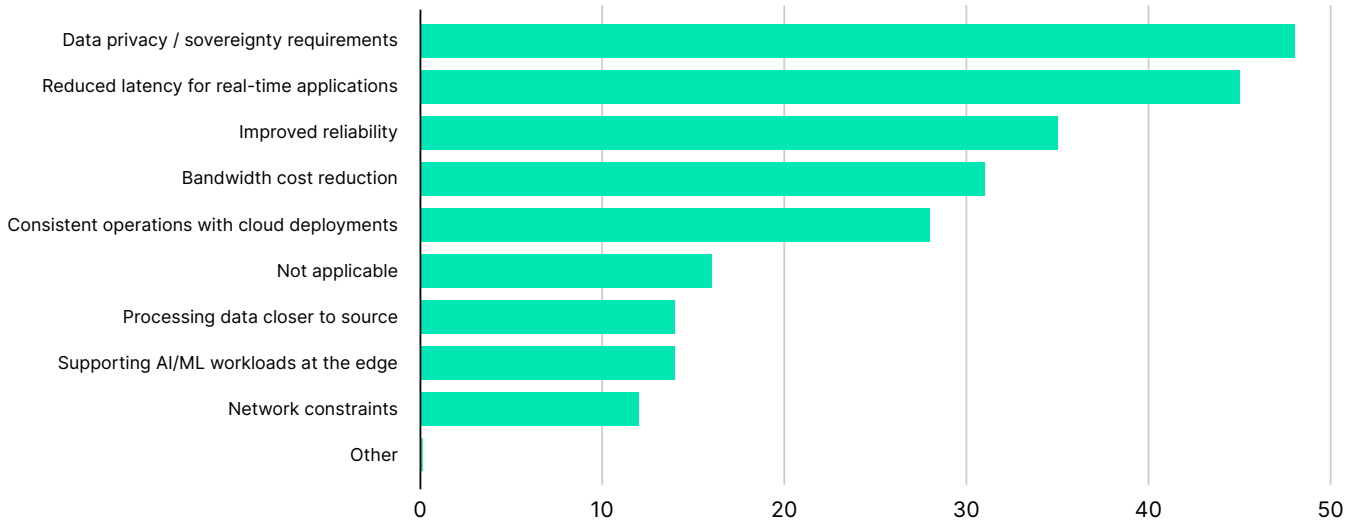
1. Edge Computing is Essential

61% say edge computing is essential for their future data strategy.

This represents a fundamental architectural shift. The drivers are clear:

Primary Drivers for Kubernetes at the Edge

If edge computing is important to you, what are the primary drivers for adopting Kubernetes at the edge? (Select top THREE)



Organizations are moving computation closer to data sources to reduce latency, respect data sovereignty requirements, and reduce bandwidth costs—all critical for AI/ML applications.

2. Real-time Processing is Critical

64% say real-time data processing is essential for their AI strategy.

The shift from batch to real-time processing reflects:

- AI applications requiring real-time responses
- Event-driven architectures
- Streaming analytics
- Real-time feature engineering for ML models

This has profound implications for data infrastructure—organizations need systems that can process, store, and serve data with millisecond latencies.

3. Training Still Dominates

58% focus primarily on training workloads vs 42% on inference.

This surprising finding contradicts the conventional wisdom that inference would dominate. The training focus reflects:

- Rapid model evolution requiring continuous retraining
- Need for domain-specific fine-tuning
- Organizations building proprietary models
- Foundation model customization

This suggests the AI/ML landscape is less mature than often portrayed—organizations are still in the model development phase rather than pure deployment/serving.

4. Vector Databases are Critical Infrastructure

77% view vector databases as critical infrastructure for AI workloads.

This is the strongest signal in the survey. The RAG architecture has become the dominant pattern for production AI applications. Vector databases enable:

- Semantic search over large document corpora
- Real-time knowledge retrieval for LLMs
- Domain-specific knowledge augmentation
- Efficient similarity search at scale

The rapid adoption of vector databases represents the fastest infrastructure shift we've tracked in DoK history.

Methodology

This report is based on a survey of 182 technology professionals conducted in October 2025. Respondents represent organizations actively running data workloads on Kubernetes.

The survey methodology included:

- **Attention check questions** to filter out low-quality responses
- **Quantitative analysis** of workload types, adoption patterns, and optimization strategies
- **Cost and performance metrics** to understand operational challenges
- **Strategic positioning questions** to identify future trends
- **Cross-reference with previous years' data** (2021, 2022, 2024) for trend analysis

Respondent Profile

- **Industries:** Technology/Software (54.40%), Financial Services (21.43%), Healthcare (3.30%)
- **Company Size:** Majority mid-to-large enterprise (19.78% at $250M-$500M revenue, 12.64% at $1B-$10B)
- **Roles:** Developer/Software Engineer (21.43%), Manager (17.58%), DevOps (10.44%), Architect (8.79%)
- **Decision-Making:** 50% are key decision-makers, 29.12% directly influence decisions

Conclusion

The 2025 Data on Kubernetes survey reveals a community at an inflection point. The adoption question has been decisively answered—DoK is now standard infrastructure for data workloads. The new question facing organizations is: "How do we achieve operational excellence?"

Four trends will define the next chapter of DoK:

Cost optimization has become the top priority. As the #1 organizational priority for 2025, cost management is no longer an afterthought but a strategic imperative. Organizations need sophisticated strategies for managing the full lifecycle of data—from ingestion through processing to long-term retention—while balancing performance requirements with budget constraints.

AI/ML workloads are driving infrastructure evolution. Vector databases have seen the fastest adoption of any infrastructure component we've tracked, while real-time data processing has become essential for AI strategies. The infrastructure that seemed cutting-edge for traditional data workloads is being re-evaluated through the lens of AI/ML requirements.

The edge + real-time architectural shift is underway. Organizations are moving away from centralized, batch-oriented systems toward distributed, real-time architectures. This shift is driven by latency requirements, data sovereignty concerns, and the need to process data closer to its source.

Performance gaps represent opportunities. While DoK has matured significantly, clear optimization opportunities remain—particularly in model loading times, storage performance, and GPU utilization. Organizations that master these optimizations will have significant competitive advantages.

The persistent skills gap remains the wild card. Organizations need practitioners who understand Kubernetes operations, data workload optimization, and AI/ML infrastructure requirements. Until this gap is addressed through training, tooling, and best practices, it will limit how quickly organizations can move from adoption to excellence.

The Data on Kubernetes Community continues to grow and evolve, providing resources, best practices, and forums for knowledge sharing among practitioners. Through regular meetups, workshops, and online events, we foster collaboration and innovation in the DoK ecosystem.

The future is being built now, and it's being built on Kubernetes.

About Data on Kubernetes Community

The Data on Kubernetes Community (DoKC) is an openly governed group of practitioners sharing in the emergence and development of techniques for the use of Kubernetes for data. Founded in June 2020, DoKC exists to assist in the emergence and development of techniques for the use of Kubernetes for data.

Our community continues to grow and evolve, providing resources, best practices, and forums for knowledge sharing among practitioners. Through regular meetups, workshops, and online events, we foster collaboration and innovation in the DoK ecosystem.

https://dok.community

DoK Community