gruve

SOLUTION BRIEF

OpenShift AI Design Service: Transform complexity into deployment confidence

01

Business problem

Organizations understand they need enterprise AI platforms, but translating requirements into production-ready architecture remains a critical barrier. Generic reference architectures from vendors provide starting points but don't address specific infrastructure constraints, compliance requirements, integration challenges, or workload characteristics that make each environment unique.

The consequences of inadequate design are severe and expensive. Attempting to design during implementation forces costly mid-project pivots when teams discover architectural flaws. Poor architecture decisions lead to $500,000-$2,000,000+ in rework expenses as organizations rebuild incorrectly designed systems. Over-engineered solutions waste 30-40% of infrastructure budgets on unnecessary capabilities, while under-designed systems create performance bottlenecks preventing production deployment.

Security and compliance gaps discovered late in implementation cause 3-6 month delays as teams retrofit controls and documentation. Ambiguous specifications lead to miscommunication between stakeholders, with platform teams building systems that don't meet data science team needs. Failed architectures damage organizational confidence, creating reluctance to approve future AI investments.

Organizations need production-ready architecture specifications that eliminate ambiguity, prevent costly mistakes, and accelerate deployment timelines—delivered by experts who have designed and implemented dozens of enterprise AI platforms.

30-40%

of infrastructure budget wasted on unnecessary capabilities

02

Why now

The gap between AI platform selection and successful deployment is where most organizations struggle. Red Hat OpenShift AI provides powerful capabilities, but architecting those capabilities for enterprise production use requires specialized expertise combining infrastructure design, AI practices, security hardening, and operational considerations.

Attempting DIY architecture design extends deployment timelines by 4-6 months as teams research best practices, evaluate design options, and revise initial approaches based on lessons learned. This trial-and-error approach wastes time while competitors deploy functioning platforms and gain market advantages.

The regulatory environment around AI systems is solidifying. Organizations must architect proper governance controls, model explainability, audit trails, and compliance capabilities from the beginning. Retrofitting these requirements into poorly designed systems costs exponentially more than incorporating them during initial architecture.

Cloud infrastructure costs for AI workloads are substantial—hundreds of thousands to millions annually. Proper architecture design optimizes spending through right-sized resource allocation, efficient workload placement, and cost-effective storage strategies. Poor design decisions create permanent cost overruns that compound yearly.

AI technology platforms evolve rapidly. Architecture designed today must accommodate future enhancements without requiring complete rebuilds. Expert design incorporates flexibility and growth paths that DIY efforts often miss, leading to architectural dead-ends requiring costly platform replacements.

03

Solution overview

Gruve's OpenShift AI Design Service delivers comprehensive, production-ready reference architectures for deploying Red Hat OpenShift AI at enterprise scale. Our certified architects design customized solutions addressing your specific infrastructure constraints, compliance requirements, and ML workload characteristics—providing implementation-ready specifications that accelerate deployment by 4-6 months and prevent costly redesigns.

Unlike vendor-provided generic templates requiring extensive customization, Gruve delivers detailed specifications tailored to your environment. Our architects combine Red Hat best practices with real-world implementation experience from dozens of enterprise deployments to create architectures optimized for your specific needs—not theoretical ideal states.

Gruve's solution components	Description
Infrastructure architecture	Compute node specifications with CPU/GPU allocation strategies, storage architecture spanning block/file/object storage, network design including load balancing and service mesh, hybrid cloud connectivity patterns, and capacity planning with growth projections
OpenShift AI platform design	Workbench configurations for data science teams, data science pipeline architecture, model registry design, model serving strategies (batch/real-time/edge), monitoring and observability framework, and resource management policies
Security & compliance architecture	Identity and access management integration, network segmentation and micro-segmentation, encryption strategy for data at rest and in transit, secrets management, audit logging, and compliance controls mapped to regulatory requirements
AI workflow design	CI/CD pipeline architecture, automated testing frameworks, model validation workflows, deployment automation, rollback procedures, A/B testing capabilities, and experiment tracking integration
Data pipeline architecture	Feature store design, data versioning strategies, ETL/ELT pipeline integration, data quality monitoring, data governance controls, and data catalog integration
Operational design	Monitoring and alerting strategy, backup and recovery procedures, disaster recovery architecture, capacity management processes, performance optimization guidance, and cost management framework

04

Benefits of Gruve's solution

Accelerated deployment timeline
Reduce time from design completion to production deployment by 4-6 months through detailed, implementation-ready specifications. Organizations report 40-50% faster deployment compared to DIY design approaches requiring extensive research and iteration.

Prevented costly redesign
Avoid $500K-$2M+ in rework expenses by getting architecture right the first time. Expert design prevents performance bottlenecks, security gaps, and integration failures that force mid-implementation pivots and costly corrections.

Optimized infrastructure investment
Achieve 25-40% reduction in infrastructure costs through right-sized compute, GPU, and storage investments based on actual workload requirements. Eliminate over-provisioning while ensuring performance meets business SLAs.

Compliance & security confidence
Design security controls and compliance capabilities from the beginning, preventing 3-6 month delays from security review failures. Meet regulatory requirements (HIPAA, GDPR, SOC 2, PCI-DSS) with audit-ready architecture documentation.

Team enablement
Provide clear, detailed specifications that eliminate ambiguity and decision fatigue for implementation teams. Reduce deployment issues by 50% and accelerate problem resolution by 60% through comprehensive documentation and design rationale

05

Service offerings

Tier	Foundation design	Comprehensive design
Duration	3-4 weeks	5-6 weeks
Architecture documentation	20-30 pages	100+ pages
Infrastructure design	Complete	Multi-cluster advanced
Security architecture	Standard	Zero-trust design
AIOps design	Basic workflows	Advanced automation
Disaster recovery	High-level	Detailed procedures
Compliance mapping	Basic	Comprehensive
Cost modeling	High-level	Detailed TCO analysis
Edge architecture	—	If required
IaC templates	Guide only	Detailed specifications
Operational runbooks	—	Complete procedures
Customer time required	15-25 hours	30-50 hours
Best for	Standard environments, straightforward requirements	Complex/multi-cloud, heavy compliance needs

06

Use case/case study

Before Gruve's OpenShift AI Design Service:

Organizations attempt to translate vendor reference architectures into their specific environments through internal design efforts. Platform architects spend weeks researching best practices, reviewing documentation, and evaluating design alternatives without clear guidance on trade-offs. Each design decision requires extensive discussion and debate among stakeholders with different priorities.

Security teams raise concerns about designs that don't meet compliance requirements, forcing architecture revision. Performance assumptions prove incorrect when workload testing reveals bottlenecks. Integration with existing systems proves more complex than anticipated, requiring design changes. The iterative cycle of design-review-revise extends timelines by months.

When organizations finally move to implementation, ambiguous specifications cause confusion and delays. Implementation teams make their own interpretation of incomplete designs, leading to systems that don't match stakeholder expectations. Missed requirements surface during testing, forcing expensive rework. Leadership loses confidence as timelines slip and costs escalate.

After Gruve's OpenShift AI Design Service:

Organizations receive comprehensive, implementation-ready architecture specifications addressing all aspects of OpenShift AI deployment. Detailed design documents eliminate ambiguity around infrastructure sizing, security controls, integration patterns, and operational procedures. Network diagrams provide exact specifications for implementation teams.

Security and compliance requirements are architected from the beginning with controls mapped to specific regulatory requirements. Performance characteristics are validated against workload requirements with capacity planning projections. Integration with existing systems is designed in detail with specific API patterns and data flows documented.

Implementation teams receive clear specifications enabling confident execution without interpretation or guesswork. Stakeholders align on architecture before implementation begins, eliminating costly mid-project debates. Risk mitigation strategies address potential implementation challenges proactively. Organizations move forward with validated designs that accelerate deployment while preventing expensive mistakes.

07

Schedule your OpenShift AI Design Consultation

Transform architectural complexity into implementation confidence with Gruve's expert design services. Our Red Hat certified architects deliver production-ready specifications optimized for your environment, requirements, and constraints.

Design engagements typically begin within 2-3 weeks of readiness assessment completion. Contact us to discuss your architecture requirements and determine which service tier best matches your needs.

Website: https://www.gruve.ai/

Email: info@gruve.ai