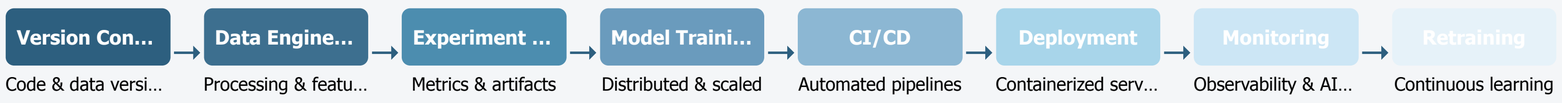
Complete AI System Pipeline

End-to-End MLOps Architecture for Production AI

This comprehensive MLOps architecture integrates `version control`, `data management`, `model training`, `deployment`, and `monitoring` to deliver scalable AI applications with reliability, reproducibility, and continuous improvement.



Code & Data Management

Version Control & Code Management

- **Git:** Code repository for tracking changes
- **GitHub/GitLab/Bitbucket:** Collaborative development
- **Code Reviews:** Quality assurance workflows
- **Branching Strategies:** Feature/release management

`Git` `GitHub` `GitLab` `Bitbucket`

Data Versioning & Management

- **DVC:** Data Version Control for datasets
- **LakeFS/Pachyderm:** Data lake versioning
- **Great Expectations:** Data validation
- **ETL Pipelines:** Data preparation & transformation

`DVC` `LakeFS` `Pachyderm` `Great Expectations`

Data Storage & Feature Management

- **S3/GCS/Azure Blob:** Cloud object storage
- **Delta Lake/Iceberg:** Data lake technology
- **Feature Stores:** Feature management & serving
- **Data Catalogs:** Metadata management

`AWS S3` `Delta Lake` `Feast` `Tecton` `Databricks`

Experiment Tracking & Management

- **MLflow:** Track experiments & metrics
- **Weights & Biases:** Visualization & collaboration
- **ClearML/Neptune:** Experiment management
- **Hyperparameter Optimization:** Search & tuning

`MLflow` `Weights & Biases` `ClearML` `Neptune.ai` `Optuna`

Training & Infrastructure

Compute Infrastructure & Scaling

- **Horizontal Scaling:** Distribute across multiple machines
- **Vertical Scaling:** More powerful resources (GPUs, TPUs)
- **Kubernetes (K8s):** Container orchestration
- **Auto Scaling Groups:** Dynamic resource allocation

`Kubernetes (K8s)` `AWS ASG` `GKE` `AKS` `KNative`

Distributed Training & Optimization

- **Ray:** Distributed compute framework
- **Horovod:** Distributed deep learning
- **PyTorch/TensorFlow:** Training frameworks
- **GPU/TPU Acceleration:** Hardware optimization

`Ray` `Horovod` `Spark MLlib` `NVIDIA RAPIDS` `DeepSpeed`

Cloud Infrastructure Services

AWS:
- SageMaker
- EC2/EKS
- Lambda
- Step Functions

Google Cloud:
- Vertex AI
- GKE
- TPUs
- Cloud Run

Azure:
- Machine Learning
- AKS
- Functions
- Databricks

Containerization & Virtualization

- **Docker:** Containerization for AI workloads
- **Podman/Singularity:** Alternative containers
- **Container Registries:** Docker Hub, ECR, GCR
- **VM Orchestration:** OpenStack, VMware

`Docker` `Podman` `Singularity` `Apptainer` `ECR`

CI/CD & Deployment

CI/CD & Workflow Automation

- **GitHub Actions/GitLab CI:** Pipeline automation
- **Jenkins/Tekton:** Custom CI/CD pipelines
- **Argo Workflows:** Kubernetes-native workflows
- **Kubeflow Pipelines:** ML workflow automation

`GitHub Actions` `GitLab CI/CD` `Jenkins` `Argo Workflows` `Kubeflow Pipelines`

Model Packaging & Deployment

- **TensorFlow Serving:** TensorFlow model serving
- **TorchServe:** PyTorch model serving
- **ONNX Runtime:** Cross-framework inference
- **KServe/Seldon Core:** Kubernetes model serving

`TensorFlow Serving` `TorchServe` `ONNX Runtime` `KServe` `Seldon Core` `BentoML`

Serverless & API Deployment

- **AWS Lambda/Azure Functions:** Serverless compute
- **FastAPI/Flask:** API frameworks for model serving
- **API Gateways:** Request routing & management
- **Cloud Run/Fargate:** Managed container deployment

`AWS Lambda` `Cloud Functions` `FastAPI` `API Gateway` `Cloud Run`

Deployment Strategies & Patterns

- **Blue-Green Deployment:** Zero-downtime updates
- **Canary Releases:** Gradual rollout of models
- **Shadow Mode:** Test models in production
- **A/B Testing:** Model comparison in production

`Istio` `Argo Rollouts` `Spinnaker` `Flagger`

Monitoring, Observability & Governance

Model Monitoring & Observability

- **Prometheus/Grafana:** Metrics & dashboards
- **Evidently AI:** Data & model drift detection
- **WhyLabs/Arize:** ML monitoring & observability
- **Distributed Tracing:** OpenTelemetry, Jaeger

`Prometheus` `Grafana` `Evidently AI` `WhyLabs` `Arize AI`

AIOps & System Monitoring

- **ELK Stack:** Log aggregation & analysis
- **Datadog/New Relic:** Application performance monitoring
- **CloudWatch/Stackdriver:** Cloud monitoring
- **Anomaly Detection:** System health monitoring

`ELK Stack` `Datadog` `New Relic` `CloudWatch` `OpenTelemetry`

Retraining & Continuous Learning

- **Automated Retraining:** Scheduled & trigger-based
- **Airflow/Prefect:** Workflow orchestration
- **Online Learning:** Continuous model updates
- **Model Evaluation:** Performance assessment

`Airflow` `Prefect` `Flyte` `MLflow` `H2O.ai`

Model Governance & Security

- **AI Explainability:** Interpretability & fairness
- **Compliance:** Regulatory & ethical frameworks
- **MLflow Model Registry:** Versioning & governance
- **IAM & RBAC:** Access control for models

`AI Explainability 360` `Fairlearn` `MLflow Model Registry` `AWS IAM` `SecML`

Deployment Architectures Comparison

Architecture	Best For	Scalability	Complexity	Key Technologies
Kubernetes-based	Enterprise-grade ML with high availability	Excellent	High	K8s, KServe, Istio, Seldon Core
Serverless	Variable workloads, cost optimization	Very Good	Medium	Lambda, Cloud Functions, Cloud Run
Managed ML Services	Rapid deployment, minimal DevOps	Good	Low	SageMaker, Vertex AI, Azure ML
On-Premises	High security, data sovereignty	Limited	Very High	Kubeflow, Seldon, Docker Swarm
Hybrid Cloud	Balancing performance & flexibility	Very Good	High	Anthos, Azure Arc, AWS Outposts

Building Production-Grade AI with MLOps Best Practices

Scalable • Reliable • Observable • Secure • Continuously Improving