Training Course on MLOps for Real-time Inference

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

Training Course on MLOps for Real-time Inference: Optimizing Models for Low-Latency Predictions is meticulously designed to equip professionals with the cutting-edge skills and practical expertise needed to deploy, manage, and optimize machine learning models for **low-latency predictions** in production environments. In today's hyper-connected world, the demand for instant insights and immediate decision-making is paramount, driving the critical need for **high-performance AI systems**. This course dives deep into the **MLOps lifecycle**, focusing specifically on the challenges and solutions associated with achieving millisecond-level response times for **AI inference**. Participants will master **model optimization techniques**, **scalable deployment strategies**, and **robust monitoring frameworks** essential for building resilient and efficient real-time ML pipelines.

The curriculum emphasizes a **hands-on, practical approach**, integrating real-world **case studies** and **industry best practices** to solidify learning. From understanding **model serving architectures** to implementing **CI/CD for ML**, participants will gain actionable knowledge to transform their **machine learning workflows** from experimental stages to **production-grade, high-throughput systems**. This training is crucial for organizations aiming to unlock the full potential of their data by enabling **real-time AI applications** across various sectors, ensuring competitive advantage through **predictive analytics** and **intelligent automation**.

Programme Curriculum

Training Course on MLOps for Real-time Inference: Optimizing Models for Low-Latency Predictions

Introduction

Training Course on MLOps for Real-time Inference: Optimizing Models for Low-Latency Predictions is meticulously designed to equip professionals with the cutting-edge skills and practical expertise needed to deploy, manage, and optimize machine learning models for **low-latency predictions** in production environments. In today's hyper-connected world, the demand for instant

insights and immediate decision-making is paramount, driving the critical need for **high-performance AI systems**. This course dives deep into the **MLOps lifecycle**, focusing specifically on the challenges and solutions associated with achieving millisecond-level response times for **AI inference**. Participants will master **model optimization techniques**, **scalable deployment strategies**, and **robust monitoring frameworks** essential for building resilient and efficient real-time ML pipelines.

The curriculum emphasizes a **hands-on, practical approach**, integrating real-world **case studies** and **industry best practices** to solidify learning. From understanding **model serving architectures** to implementing **CI/CD for ML**, participants will gain actionable knowledge to transform their **machine learning workflows** from experimental stages to **production-grade, high-throughput systems**. This training is crucial for organizations aiming to unlock the full potential of their data by enabling **real-time AI applications** across various sectors, ensuring competitive advantage through **predictive analytics** and **intelligent automation**.

Course Duration

10 days

Course Objectives

1. Comprehend the core principles and MLOps lifecycle for productionizing ML models.
2. Implement advanced model optimization techniques like quantization, pruning, and knowledge distillation for fast inference.
3. architect robust and scalable model serving architectures using tools like TensorFlow Serving, NVIDIA Triton, and TorchServe.
4. Develop automated CI/CD pipelines for continuous integration, continuous delivery, and continuous training of ML models.
5. Leverage feature stores and streaming data platforms (e.g., Kafka, Flink) for consistent and low-latency feature delivery.
6. Understand and apply edge AI deployment strategies and cloud-native MLOps platforms (AWS SageMaker, GCP Vertex AI, Azure ML).
7. Establish comprehensive real-time monitoring and alerting systems for model drift, data quality, and prediction latency using Prometheus and Grafana.
8. Implement robust model versioning and experiment tracking using tools like MLflow and DVC.
9. Design and implement scalable inference solutions using Kubernetes, autoscaling, and distributed serving.
10. Conduct effective A/B testing and canary deployments for safe and controlled model rollouts.
11. Understand and apply security best practices and data governance principles for production ML systems.
12. Develop effective strategies for diagnosing and resolving issues in low-latency ML inference environments.
13. Explore LLMOps concepts and strategies for deploying and optimizing large language models for real-time applications.

Organizational Benefits

- Significantly reduce the time from model development to production deployment, enabling faster delivery of AI-powered features.
- Implement robust monitoring and automation to ensure consistent and high-quality model performance in real-world scenarios.

- Automate repetitive MLOps tasks, freeing up data scientists and engineers to focus on innovation and model improvement.
- Build adaptable and elastic systems capable of handling increasing inference loads and diverse real-time application demands.
- Optimize resource utilization and minimize manual interventions through efficient MLOps practices.
- Drive immediate value from machine learning by enabling real-time decision-making and intelligent automation.
- Foster seamless collaboration between data science, ML engineering, and operations teams.
- Proactively detect and address model degradation, data drift, and other production issues before they impact business.

Target Audience

1. Machine Learning Engineers
2. Data Scientists
3. DevOps Engineers
4. AI Architects.
5. Software Engineers.
6. Cloud Engineers.
7. Technical Leads & Managers
8. Researchers & Academics.

Course Outline

Module 1: Introduction to MLOps for Real-time Inference

- Defining MLOps in the context of low-latency predictions.
- Challenges of deploying ML models for real-time inference.
- Understanding the MLOps lifecycle from experimentation to production.
- Key differences between traditional software DevOps and MLOps.
- Overview of the course structure and learning objectives.
- **Case Study:** Early challenges faced by ride-sharing platforms in providing real-time ETA predictions due to lack of MLOps.

Module 2: Model Optimization Techniques for Low Latency

- Introduction to model quantization (e.g., INT8, FP16).
- Model pruning and sparsity techniques to reduce model size.
- Knowledge distillation: transferring knowledge from a large model to a smaller one.
- Techniques for optimizing deep learning models for inference (e.g., graph optimization).
- Benchmarking model inference speed and resource utilization.
- **Case Study:** Optimizing a fraud detection model using quantization for sub-millisecond response times in financial transactions.

Module 3: Real-time Feature Engineering and Feature Stores

- Understanding the need for real-time features in low-latency predictions.
- Designing streaming data pipelines for feature generation (Kafka, Flink).
- Introduction to Feature Stores: online and offline serving.
- Consistency challenges between training and serving features.
- Implementing feature versioning and management.
- **Case Study:** Building a real-time recommendation system using a feature store to deliver personalized product suggestions instantly.

Module 4: Model Serving Architectures

- Overview of common model serving patterns (REST APIs, gRPC).
- Specialized inference servers: TensorFlow Serving, NVIDIA Triton Inference Server, TorchServe.
- Microservices architecture for scalable model serving.
- Containerization with Docker for consistent deployment environments.
- Serverless inference options (AWS Lambda, Azure Functions).
- **Case Study:** Deploying a computer vision model for real-time object detection using NVIDIA Triton Inference Server on edge devices.

Module 5: Kubernetes for Scalable Inference

- Introduction to Kubernetes for orchestrating ML workloads.
- Deploying and managing model serving deployments on Kubernetes.
- Scaling strategies: Horizontal Pod Autoscaler (HPA) for inference.
- Resource management and optimization for ML inference pods.
- Kubernetes best practices for high-availability and fault tolerance.
- **Case Study:** Scaling a natural language processing (NLP) inference service on Kubernetes to handle millions of requests per second for a chatbot.

Module 6: CI/CD for Machine Learning (CI/CD/CT)

- Adapting CI/CD principles for machine learning workflows.
- Automating model training, testing, and deployment.
- Continuous Training (CT): Triggering retraining based on performance or data drift.
- Tools for ML CI/CD: Jenkins, GitHub Actions, GitLab CI/CD.
- Orchestration tools for ML pipelines (Kubeflow Pipelines, Apache Airflow).
- **Case Study:** Implementing a CI/CD pipeline for a predictive maintenance model, enabling automatic retraining and deployment upon new data arrival.

Module 7: Model Monitoring and Alerting in Production

- Importance of real-time monitoring for production ML models.
- Detecting model drift (concept drift, data drift) and performance degradation.
- Monitoring tools: Prometheus, Grafana, Evidently AI.
- Setting up alerts for anomalies and critical issues.
- Logging and observability for ML inference services.
- **Case Study:** Detecting and alerting on unusual credit card transaction patterns using real-time model monitoring to prevent fraud.

Module 8: Model Governance and Reproducibility

- Establishing model registries and version control.
- Tracking experiments, parameters, and metrics with MLflow.
- Data versioning with DVC for reproducible experiments.
- Ensuring traceability and auditability of ML models.
- Compliance and regulatory considerations for AI systems.
- **Case Study:** A pharmaceutical company ensuring strict reproducibility of drug discovery ML models for regulatory compliance.

Module 9: A/B Testing and Canary Deployments

- Strategies for deploying new model versions safely.

- Implementing A/B testing for comparing model performance in production.
- Canary deployments for gradual rollouts and risk mitigation.
- Blue/Green deployments for zero-downtime updates.
- Rollback strategies for faulty deployments.
- **Case Study:** A marketing company A/B testing different recommendation algorithms to optimize user engagement and conversion rates.

Module 10: Cloud-Native MLOps Platforms

- Overview of cloud-specific MLOps offerings (AWS SageMaker, Google Cloud Vertex AI, Azure Machine Learning).
- Leveraging managed services for scalable and reliable MLOps.
- Serverless and cost-effective inference options in the cloud.
- Integrating with cloud-native data and compute services.
- Choosing the right cloud platform for your MLOps needs.
- **Case Study:** A retail giant leveraging Google Cloud Vertex AI to manage its entire MLOps workflow for real-time inventory optimization.

Module 11: Edge AI and On-Device Inference

- Introduction to edge computing and its relevance for low-latency AI.
- Deploying optimized models directly on edge devices (IoT, mobile).
- Challenges and considerations for resource-constrained environments.
- Hardware acceleration for edge inference (TPUs, NPUs).
- Managing and updating models on a fleet of edge devices.
- **Case Study:** Smart factory using edge AI for real-time defect detection on the assembly line, minimizing latency and bandwidth usage.

Module 12: Performance Optimization for Real-time Systems

- Advanced techniques for reducing inference latency (batching, parallel processing).
- Optimizing network communication and data serialization.
- Memory management for large models in production.
- Profiling and identifying bottlenecks in the inference pipeline.
- Techniques for cold start reduction and warm-up.
- **Case Study:** Optimizing a real-time bidding system for ad placement to achieve microsecond-level prediction times.

Module 13: Security and Privacy in MLOps

- Securing ML models and inference endpoints.
- Data privacy considerations (GDPR, HIPAA, CCPA) in real-time ML.
- Access control and authentication for MLOps pipelines.
- Threat modeling for ML systems.
- Federated learning for privacy-preserving AI.
- **Case Study:** A healthcare provider implementing strict security and privacy measures for real-time patient diagnosis models.

Module 14: Troubleshooting and Debugging Production ML Systems

- Common pitfalls and issues in real-time ML deployments.
- Strategies for debugging inference failures and performance drops.
- Leveraging logs, metrics, and tracing for root cause analysis.
- Creating effective incident response plans for ML systems.

- Post-mortem analysis for continuous improvement.
- **Case Study:** Diagnosing and resolving a sudden spike in latency for an online gaming matchmaking system.

Module 15: Advanced Topics & Future Trends in MLOps

- LLMOps: Operationalizing Large Language Models for real-time use cases.
- Responsible AI and Ethical MLOps considerations.
- Data-centric AI and its impact on MLOps.
- MLOps for AIOps: Using AI to manage IT operations.
- Emerging tools and technologies in the MLOps landscape.
- **Case Study:** Discussing the challenges and solutions for deploying a real-time conversational AI assistant powered by a large language model.

Training Methodology

This course will employ a blended learning approach, combining:

- **Interactive Lectures:** Core concepts and theoretical foundations.
- **Hands-on Labs:** Practical exercises using industry-standard MLOps tools and cloud platforms.
- **Live Coding Demonstrations:** Walkthroughs of real-world MLOps implementations.
- **Case Study Analysis:** In-depth discussion of successful and challenging MLOps scenarios.
- **Group Discussions:** Collaborative problem-solving and knowledge sharing.
- **Q&A Sessions:** Direct interaction with instructors to clarify doubts.
- **Project-Based Learning:** A capstone project to apply learned concepts to a comprehensive MLOps scenario.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Online	$2,000
23 Nov 2026	04 Dec 2026	Physical	$3,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$9,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$12,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com