Training Course on Scalable ML Serving with Kubeflow/Sagemaker

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 Scalable ML Serving with Kubeflow/Sagemaker: Deploying and Scaling Models in Cloud Environments focuses on empowering data scientists, ML engineers, and DevOps professionals with the essential skills to deploy and scale machine learning models efficiently in cloud environments. Leveraging industry-leading platforms like **Kubeflow** and **Amazon SageMaker**, participants will master the intricacies of **MLOps**, **model serving**, and **distributed training**, ensuring robust, high-performance, and cost-effective ML solutions in production.

The program delves into the critical aspects of transforming experimental ML models into reliable, enterprise-grade applications. Through practical, hands-on labs and real-world case studies, attendees will gain expertise in building **end-to-end MLOps pipelines**, automating **model deployment**, implementing **scalable inference architectures**, and effectively monitoring their ML systems. This course is designed to bridge the gap between ML development and production, addressing challenges related to **resource optimization**, **latency reduction**, and **high availability** for deep learning and traditional ML models.

Programme Curriculum

Training Course on Scalable ML Serving with Kubeflow/Sagemaker: Deploying and Scaling Models in Cloud Environments

Introduction

Training Course on Scalable ML Serving with Kubeflow/Sagemaker: Deploying and Scaling Models in Cloud Environments focuses on empowering data scientists, ML engineers, and DevOps professionals with the essential skills to deploy and scale machine learning models efficiently in cloud environments. Leveraging industry-leading platforms like **Kubeflow** and **Amazon SageMaker**, participants will master the intricacies of **MLOps**, **model serving**, and **distributed training**, ensuring robust, high-performance, and cost-effective ML solutions in production.

The program delves into the critical aspects of transforming experimental ML models into reliable, enterprise-grade applications. Through practical, hands-on labs and real-world case studies, attendees will gain expertise in building **end-to-end MLOps pipelines**, automating **model deployment**, implementing **scalable inference architectures**, and effectively monitoring their ML systems. This course is designed to bridge the gap between ML development and production, addressing challenges related to **resource optimization**, **latency reduction**, and **high availability** for deep learning and traditional ML models.

Course Duration

10 days

Course Objectives

1. Understand the core components and architecture of Kubeflow for orchestrating ML workflows on Kubernetes.
2. Gain proficiency in leveraging SageMaker's managed services for various ML lifecycle stages.
3. Design and build automated, reproducible, and scalable ML pipelines.
4. Configure and manage low-latency model serving endpoints using Kubeflow Serving (KServe) and SageMaker Endpoints.
5. Learn to process large datasets efficiently using batch transform jobs on both platforms.
6. Develop strategies for efficient GPU/CPU allocation and cost management in cloud ML deployments.
7. Implement robust monitoring and alerting for model drift, data quality, and prediction latency.
8. Configure and execute distributed training jobs for large-scale deep learning models.
9. Integrate ML pipelines into existing CI/CD workflows for continuous model delivery.
10. Establish best practices for tracking model versions and their associated metadata.
11. Implement security measures for ML endpoints, data, and access control.
12. Diagnose and resolve challenges related to model serving and scaling in production.
13. Understand how to combine Kubeflow and SageMaker for flexible and powerful ML solutions.

Organizational Benefits

- Rapidly deploy and iterate on machine learning models, bringing AI-powered solutions to users faster.
- Automate repetitive MLOps tasks, reducing manual effort and potential errors.
- Optimize resource allocation and leverage cost-effective deployment strategies.
- Ensure high availability, low latency, and consistent performance of deployed ML models.
- Enable continuous feedback loops and proactive model management for better business outcomes.
- Build robust ML infrastructure capable of handling growing data volumes and user demands.
- Foster an internal culture of MLOps excellence, leading to more innovative and reliable AI products.

Target Audience

1. Machine Learning Engineers
2. Data Scientists.
3. DevOps Engineers.
4. Cloud Architects.

5. Software Engineers.
6. AI/ML Leads & Managers
7. Solution Architects.
8. Data Engineers

Course Outline

Module 1: Introduction to Scalable ML Serving & MLOps

- Concepts: Overview of MLOps principles, challenges in productionizing ML models, and the importance of scalability.
- Platforms: Introduction to Kubeflow and Amazon SageMaker as leading MLOps platforms.
- Architecture: Understanding common architectural patterns for scalable ML serving.
- Benefits: Discussing the business impact of robust ML deployment strategies.
- **Case Study:** Analyzing a startup's journey from prototype to production with early ML models and the challenges faced.

Module 2: Kubernetes Fundamentals for ML

- Concepts: Core Kubernetes concepts: Pods, Deployments, Services, Namespaces.
- Orchestration: How Kubernetes orchestrates containerized applications.
- Resource Management: CPU, memory, and GPU resource requests and limits.
- Networking: Understanding Kubernetes networking for ML applications.
- **Case Study:** Deploying a simple Flask API for model inference on a Kubernetes cluster.

Module 3: Introduction to Kubeflow

- Components: Deep dive into Kubeflow components: Pipelines, KFServing (KServe), Notebooks, Katib, etc.
- Installation: Overview of deploying Kubeflow on Kubernetes.
- Dashboard: Navigating the Kubeflow Central Dashboard.
- User Profiles: Managing multi-tenancy in Kubeflow.
- **Case Study:** Setting up a new Kubeflow environment and exploring its functionalities.

Module 4: Kubeflow Pipelines for ML Workflows

- Concepts: Defining and orchestrating end-to-end ML workflows using Kubeflow Pipelines.
- Components: Building reusable pipeline components.
- SDK: Using the Kubeflow Pipelines SDK for authoring pipelines in Python.
- Execution: Running and monitoring pipeline runs.
- **Case Study:** Building a multi-step ML pipeline for data preprocessing, training, and model registration.

Module 5: Model Serving with Kubeflow Serving (KServe)

- Concepts: Introduction to KServe for model inference on Kubernetes.
- Inferencing: Deploying various model types (TensorFlow, PyTorch, Scikit-learn, XGBoost).
- Traffic Management: Canary deployments, A/B testing, and traffic splitting.
- Autoscaling: Configuring horizontal pod autoscaling for inference endpoints.
- **Case Study:** Deploying a deep learning image classification model with KServe and testing its scalability under load.

Module 6: Advanced Kubeflow Techniques

- Distributed Training: Leveraging Kubeflow operators for distributed TensorFlow, PyTorch, etc.
- Hyperparameter Tuning: Using Katib for automated hyperparameter optimization.
- Metadata Management: Tracking model artifacts and lineage with ML Metadata (MLMD).
- Custom Resources: Extending Kubeflow with custom components and operators.
- **Case Study:** Optimizing a fraud detection model using Katib and tracking experiment results with MLMD.

Module 7: Introduction to Amazon SageMaker

- Overview: Exploring SageMaker's fully managed ML service offerings.
- SageMaker Studio: Navigating the integrated development environment.
- Key Services: Understanding SageMaker Notebook Instances, Training, Endpoints, and Pipelines.
- IAM Roles: Managing permissions and security in SageMaker.
- **Case Study:** Setting up a SageMaker Studio environment and running a basic notebook.

Module 8: Model Training with Amazon SageMaker

- Algorithms: Utilizing built-in SageMaker algorithms and custom containers.
- Distributed Training: Scaling training jobs using SageMaker's distributed capabilities.
- Hyperparameter Tuning: Automating hyperparameter optimization with SageMaker HPO.
- Spot Instances: Cost optimization for training jobs.
- **Case Study:** Training a large language model (LLM) on SageMaker using distributed training and HPO.

Module 9: Model Deployment with Amazon SageMaker Endpoints

- Real-time Inference: Deploying models to real-time SageMaker Endpoints.
- Endpoint Configuration: Managing instance types, autoscaling, and endpoint variants.
- Batch Transform: Performing offline inference on large datasets.
- Model Monitoring: Setting up SageMaker Model Monitor for drift detection.
- **Case Study:** Deploying a personalized recommendation model to a SageMaker endpoint and configuring autoscaling.

Module 10: Building End-to-End MLOps with SageMaker Pipelines

- Concepts: Orchestrating multi-step ML workflows using SageMaker Pipelines.
- Steps: Defining pipeline steps for data processing, training, model registration, and deployment.
- SDK: Authoring pipelines with the SageMaker Python SDK.
- CI/CD Integration: Automating pipeline execution with AWS CodePipeline/CodeBuild.
- **Case Study:** Implementing a complete MLOps pipeline for a predictive maintenance model.

Module 11: Hybrid Architectures: Kubeflow and SageMaker Integration

- Use Cases: Identifying scenarios where combining Kubeflow and SageMaker offers advantages.
- SageMaker Components for Kubeflow Pipelines: Integrating SageMaker jobs into Kubeflow workflows.
- Data Exchange: Strategies for seamless data transfer between environments.
- Best Practices: Guidelines for designing and managing hybrid ML solutions.
- **Case Study:** Building a hybrid architecture where data preprocessing and initial model training happen on Kubeflow, while large-scale training and deployment are managed by

SageMaker.

Module 12: Monitoring and Observability for ML Systems

- Metrics: Key metrics for monitoring ML model performance (accuracy, latency, throughput).
- Logging: Centralized logging for ML applications in production.
- Alerting: Setting up alerts for anomalies and performance degradation.
- Tools: Using Prometheus, Grafana, CloudWatch, and custom dashboards.
- **Case Study:** Implementing a comprehensive monitoring solution for a deployed ML model, including custom dashboards and alert configurations.

Module 13: Securing ML Deployments

- Access Control: Implementing IAM roles and Kubernetes RBAC for ML resources.
- Network Security: Securing endpoints and communication channels.
- Data Encryption: Protecting data at rest and in transit.
- Vulnerability Management: Scanning container images for security vulnerabilities.
- **Case Study:** Reviewing and hardening the security posture of a production ML system.

Module 14: Cost Optimization in Cloud ML Environments

- Instance Selection: Choosing optimal compute instances for training and inference.
- Autoscaling Strategies: Implementing cost-effective autoscaling.
- Spot Instances/Managed Spot Training: Leveraging cheaper compute options.
- Resource Quotas: Managing resource consumption within budgets.
- **Case Study:** Analyzing a real-world ML deployment and identifying opportunities for cost reduction.

Module 15: Future Trends in Scalable ML Serving

- MLOps Platforms: Evolution of MLOps tools and platforms.
- Responsible AI: Explainability (XAI), fairness, and bias in production models.
- Edge AI: Deploying ML models to edge devices.
- Foundation Models: Serving and fine-tuning large foundation models.
- Ethical Considerations: Discussing the broader societal impact of scaled AI.
- **Case Study:** Exploring emerging technologies like serverless inference with AWS Lambda or Google Cloud Run for niche ML serving needs.

Training Methodology

This course employs a participatory and hands-on approach to ensure practical learning, including:

- Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
21 Sep 2026	02 Oct 2026	Online	$2,000
21 Sep 2026	02 Oct 2026	Physical	$3,000
21 Sep 2026	02 Oct 2026	Physical	$0

Start Date	End Date	Location	Price
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0

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

Register Online Now

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