Training Course on Cost Optimization in MLOps

Professional Development Training Course Outline

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

Course Introduction

In today's competitive landscape, **Machine Learning Operations (MLOps)** has become critical for successful **AI/ML model deployment** and lifecycle management. However, the rapidly escalating **cloud infrastructure** and **compute costs** associated with training, deploying, and maintaining **ML workloads** pose significant challenges for organizations. This course addresses this pressing need by providing comprehensive strategies and practical techniques for achieving **cost efficiency** within MLOps environments. We will explore how to **optimize resource utilization**, leverage **cloud-native services**, and implement **intelligent cost management** practices to unlock substantial savings without compromising model performance or development velocity.

Training Course on Cost Optimization in MLOps: Managing cloud infrastructure and compute costs for ML workloads delves deep into the intersection of **MLOps best practices** and **cloud financial management**. Participants will gain actionable insights into identifying **cost inefficiencies**, implementing **governance policies**, and adopting a **FinOps for ML** mindset. By mastering these principles, organizations can transform their ML initiatives into truly **scalable, sustainable, and profitable** ventures, driving innovation while maintaining strict budgetary control.

Programme Curriculum

Training Course on Cost Optimization in MLOps: Managing cloud infrastructure and compute costs for ML workloads

Introduction

In today's competitive landscape, **Machine Learning Operations (MLOps)** has become critical for successful **AI/ML model deployment** and lifecycle management. However, the rapidly escalating **cloud infrastructure** and **compute costs** associated with training, deploying, and maintaining **ML workloads** pose significant challenges for organizations. This course addresses this pressing need by providing comprehensive strategies and practical techniques for achieving **cost efficiency** within MLOps environments. We will explore how to **optimize resource utilization**, leverage **cloud-**

native services, and implement **intelligent cost management** practices to unlock substantial savings without compromising model performance or development velocity.

Training Course on Cost Optimization in MLOps: Managing cloud infrastructure and compute costs for ML workloads delves deep into the intersection of **MLOps best practices** and **cloud financial management**. Participants will gain actionable insights into identifying **cost inefficiencies**, implementing **governance policies**, and adopting a **FinOps for ML** mindset. By mastering these principles, organizations can transform their ML initiatives into truly **scalable, sustainable, and profitable** ventures, driving innovation while maintaining strict budgetary control.

Course Duration

10 days

Course Objectives

1. Master Cloud Cost Management principles specifically tailored for ML workloads.
2. Implement Resource Optimization strategies for GPU computing and distributed training.
3. Leverage Cloud-Native MLOps Platforms for efficient cost control.
4. Apply FinOps for AI/ML methodologies to track and reduce spending.
5. Optimize Data Storage Costs in ML data pipelines and feature stores.
6. Understand and utilize Spot Instances and Preemptible VMs for cost-effective training and inference.
7. Develop strategies for Model Serving Cost Optimization and inference efficiency.
8. Implement Automated Cost Monitoring and Alerting for MLOps pipelines.
9. Learn Budgeting and Forecasting techniques for unpredictable ML expenses.
10. Explore Serverless ML and its impact on compute cost reduction.
11. Design Cost-Efficient MLOps Architectures for scalability and savings.
12. Analyze Cost-Performance Trade-offs in model development and deployment.
13. Integrate Green AI principles for sustainable and cost-aware ML operations.

Organizational Benefits

- Achieve demonstrable savings on cloud infrastructure and compute resources, directly impacting the bottom line.
- Gain granular visibility and control over ML-related cloud spending, enabling better budgeting and forecasting.
- Optimize the allocation and consumption of expensive GPU and CPU resources, reducing waste.
- Accelerate the return on investment for machine learning projects by lowering operational expenses.
- Build future-proof MLOps pipelines that are inherently cost-efficient and environmentally conscious.
- Reduce the risk of unexpected cloud bills and budget overruns associated with burgeoning ML deployments.
- Enable rapid experimentation and deployment of ML models with optimized cost structures, fostering innovation.
- Foster a shared understanding of cost implications across data science, ML engineering, and finance teams.

Target Audience

1. ML Engineers

2. Data Scientists
3. Cloud Architects
4. DevOps Engineers.
5. FinOps Practitioners
6. AI/ML Project Managers.
7. Data Engineers.
8. Technical Leads & Team Leads

Course Outline

Module 1: Introduction to MLOps and Cloud Cost Challenges

- Overview of the MLOps lifecycle and its complexities.
- Identifying key cost drivers in ML development and deployment.
- Understanding the economic impact of unmanaged ML workloads.
- Introduction to cloud pricing models for compute, storage, and networking.
- **Case Study:** Analyzing a startup's unexpected cloud bill due to unoptimized ML experiments.

Module 2: Foundational Cloud Cost Optimization Principles

- The shared responsibility model for cloud financial management.
- Rightsizing resources: Matching instance types to workload needs.
- Leveraging reserved instances, savings plans, and spot instances.
- Strategies for identifying and eliminating idle resources.
- **Case Study:** Migrating a data processing pipeline from on-demand to spot instances, demonstrating significant savings.

Module 3: Compute Cost Optimization for ML Training

- Optimizing GPU utilization for deep learning workloads.
- Distributed training strategies and their cost implications (e.g., Horovod, PyTorch DDP).
- Hyperparameter optimization (HPO) for efficient model training.
- Strategies for early stopping and training progress monitoring.
- **Case Study:** Reducing training costs by 40% through optimized distributed training and HPO on a large image classification model.

Module 4: Data Storage Cost Management in MLOps

- Tiered storage strategies for ML datasets and artifacts (e.g., S3 Intelligent-Tiering, Glacier).
- Data lifecycle management for versioned datasets and model checkpoints.
- Optimizing data transfer costs (ingress/egress) in ML pipelines.
- Leveraging data compression and deduplication techniques.
- **Case Study:** A genomics company reducing storage costs by implementing intelligent tiering for vast genomic datasets.

Module 5: MLOps Platform-Specific Cost Optimization (AWS Focus)

- Cost-saving features in Amazon SageMaker (Managed Spot Training, Multi-Model Endpoints).
- Optimizing AWS Lambda and Amazon EKS for serverless and containerized ML.
- Leveraging AWS Cost Explorer and Cost Anomaly Detection for ML spend.
- Best practices for AWS networking and data transfer costs for ML.

- **Case Study:** An e-commerce platform reducing inference costs by using SageMaker Multi-Model Endpoints and auto-scaling.

Module 6: MLOps Platform-Specific Cost Optimization (GCP Focus)

- Cost optimization with Google Cloud AI Platform and Vertex AI.
- Utilizing Google Kubernetes Engine (GKE) and Preemptible VMs for ML.
- Analyzing GCP billing reports and cost breakdowns for ML projects.
- Strategies for optimizing BigQuery and Cloud Storage for ML data.
- **Case Study:** A financial institution achieving 30% cost savings on ML training by leveraging GCP Preemptible VMs.

Module 7: MLOps Platform-Specific Cost Optimization (Azure Focus)

- Cost management in Azure Machine Learning and Azure Kubernetes Service (AKS).
- Optimizing Azure Databricks and Azure Synapse Analytics for large-scale ML.
- Monitoring Azure costs using Azure Cost Management and Azure Advisor.
- Implementing cost-aware MLOps pipelines on Azure.
- **Case Study:** A manufacturing firm optimizing predictive maintenance model deployment costs on Azure through AKS.

Module 8: Model Serving and Inference Cost Optimization

- Strategies for efficient model deployment: batch vs. real-time inference.
- Auto-scaling and load balancing for cost-effective inference.
- Model compression and quantization techniques to reduce serving costs.
- Choosing optimal inference hardware (CPU vs. GPU, specialized chips).
- **Case Study:** A content recommendation engine lowering inference costs by 25% through model compression and CPU-based serving.

Module 9: FinOps for Machine Learning

- Introduction to FinOps principles and their application to MLOps.
- Establishing cost accountability and ownership within ML teams.
- Implementing showback and chargeback models for ML infrastructure.
- Developing cost allocation tags and strategies.
- **Case Study:** A large enterprise implementing a FinOps framework to gain transparency and control over departmental ML spending.

Module 10: Automated Cost Monitoring and Governance

- Setting up real-time cost alerts and dashboards.
- Integrating cost data with MLOps monitoring tools (e.g., Prometheus, Grafana).
- Automated shutdown of idle resources and non-production environments.
- Implementing policy-driven cost governance for ML pipelines.
- **Case Study:** An anomaly detection system for cloud costs, triggered by unexpected spikes in ML resource consumption.

Module 11: CI/CD for Cost-Efficient MLOps

- Integrating cost checks into CI/CD pipelines for ML models.
- Automating resource provisioning and de-provisioning based on workload.
- Version control for infrastructure-as-code and cost configurations.
- Testing cost implications of new model deployments before production.

- **Case Study:** A continuous integration pipeline that automatically flags high-cost resource requests for review before deployment.

Module 12: Advanced Resource Scheduling and Orchestration

- Leveraging Kubernetes scheduling for optimal resource placement.
- Dynamic resource allocation for fluctuating ML workloads.
- Workload managers and schedulers (e.g., Kubeflow, Apache Airflow) for cost control.
- Optimizing container images and runtimes for reduced footprint.
- **Case Study:** A research lab optimizing their ML experiment cluster utilization using advanced Kubernetes scheduling.

Module 13: Green AI and Sustainable MLOps

- Understanding the carbon footprint of AI/ML workloads.
- Strategies for energy-efficient model training and inference.
- Choosing cloud regions with lower carbon intensity.
- The intersection of cost optimization and environmental sustainability.
- **Case Study:** A public sector project implementing Green AI principles to reduce both costs and environmental impact.

Module 14: Cost-Performance Trade-offs and Optimization Strategies

- Balancing model accuracy, latency, and throughput with cost constraints.
- Strategies for iterative cost optimization throughout the ML lifecycle.
- Decision-making frameworks for resource allocation based on business value.
- Hybrid cloud strategies for cost-effective ML.
- **Case Study:** An autonomous driving company analyzing trade-offs between model complexity and inference cost for real-time applications.

Module 15: Future Trends in MLOps Cost Optimization

- Emerging technologies for AI/ML hardware acceleration (e.g., custom silicon).
- The impact of Generative AI on cloud costs and optimization strategies.
- Serverless inference and its evolving cost models.
- Trends in FinOps and MLOps convergence.
- **Case Study:** Exploring how a leading tech company is preparing for future ML cost challenges with specialized hardware and new architectural patterns.

Training Methodology

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

- Interactive lectures and presentations.
- Group discussions and brainstorming sessions.
- Hands-on exercises using real-world datasets.
- Role-playing and scenario-based simulations.
- Analysis of case studies to bridge theory and practice.
- Peer-to-peer learning and networking.
- Expert-led Q&A sessions.
- Continuous feedback and personalized guidance.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

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

Start Date	End Date	Location	Price
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