

Google Cloud - Generative AI in Production

Download Whitepaper: Accelerate Your Modernization Efforts with a Cloud-Native Strategy
Get Your Free Copy Now

Course Number: GCP-GAIPROD

Duration: 1 days

Overview

Course Description

Traditional MLOps is a set of practices to productionize traditional ML systems for enterprise applications. Generative AI raises new challenges in managing and productionizing applications at scale. The field of generative AI operations seeks to address these new challenges. In this course, you learn about the challenges that arise when deploying and productionizing generative AI-powered applications. You learn how to secure your generative AI-powered applications. Finally, you will discuss best practices for logging and monitoring your generative AI-powered applications in production.

Skills Gained

- Understand the challenges in productionizing applications using generative AI
- Manage experimentation and evaluation for LLM-powered application
- Productionize LLM-powered applications
- Secure generative AI applications
- Implement logging and monitoring for LLM-powered applications

Who Can Benefit

Developers, DevOps engineers and machine learning engineers who wish to operationalize GenAI-based applications

Prerequisites

Completion of the "Application Development with LLMs on Google Cloud" or equivalent knowledge.

Audience

Course Details

Introduction to Generative AI in Production

- Understand generative AI operations
- Compare traditional MLOps and GenAIOps
- Analyze the components of an LLM system
- Define and compare RAG and ReAct

Generative AI Application Deployment

- Evaluate application deployment options
- Deploy, package, and version apps
- Lab: Deploying an Agentic Application on Cloud Run

Productionizing Generative AI

- Maintain and update LLM models
- Test and evaluate gen AI-powered apps
- Deploy CI/CD pipelines for gen AI-powered apps
- Lab: Tracking Versions of Generative AI Applications

Securing Generative AI Applications

- Identify security challenges for gen AI applications
- Understand prompt security issues
- Apply sensitive data protection and DLP API
- Implement Model Armor
- Lab: Securing Generative AI-Powered Applications

Observability for Production LLM Systems

- Describe the purpose and capabilities of Google Cloud Observability
- Explain the purpose of Cloud Monitoring

- Explain the purpose of Cloud Logging
- Explain the purpose of Cloud Trace
- Lab: Logging, Monitoring, and Agent Analytics