Deploy and Manage Generative AI Models on Google Cloud

Course Duration: 24 Hours **Course code: DMGAMGC**

1. Course Overview

This course focuses on deploying, managing, and scaling Generative AI models using Google Cloud services. Learners will explore Vertex AI, foundation models, model tuning, MLOps practices, and responsible AI deployment strategies. The course also covers real-world implementation of GenAI solutions across enterprise environments.

2. What you'll learn?

By the end of the course, you will be able to:

- Understand Generative AI concepts and Google Cloud AI ecosystem
- Deploy foundation models using Vertex AI
- Fine-tune and customize large language models (LLMs)
- Build and manage end-to-end ML pipelines
- Integrate Generative AI into applications using APIs
- Implement prompt engineering techniques
- Monitor, optimize, and scale AI models
- Apply responsible AI and governance practices

3. Target Audience

- Cloud Engineers
- AI/ML Engineers
- Data Scientists
- Software Developers
- DevOps Engineers

- Solution Architects

4. Pre-Requisites

Before taking this course, you should have:

- Basic knowledge of Google Cloud Platform (GCP)
- Understanding of Machine Learning concepts
- Familiarity with Python programming
- Basic knowledge of APIs and cloud deployment

5. Course content

Module 1: Course Introduction

- Course objectives and structure
- Overview of Generative AI trends
- Introduction to Google Cloud AI services

Module 2: Introduction to Generative AI

- What is Generative AI
- Types of generative models (LLMs, Diffusion Models)
- Use cases and industry applications
- Challenges and limitations

Module 3: Google Cloud AI Ecosystem

- Overview of Vertex AI
- Google foundation models (PaLM, Gemini)
- AI infrastructure on GCP
- Integration with BigQuery and Dataflow

Module 4: Working with Vertex AI

- Navigating Vertex AI console
- Model Garden overview

- Deploying pre-trained models
- Using notebooks and workbench

Module 5: Prompt Engineering

- Principles of prompt design
- Zero-shot and few-shot prompting
- Prompt optimization techniques
- Testing and evaluating prompts

Module 6: Model Deployment

- Deploying models using Vertex AI endpoints
- Batch vs real-time predictions
- API integration for applications
- Managing model versions

Module 7: Fine-Tuning and Customization

- Introduction to model tuning
- Fine-tuning LLMs on custom datasets
- Transfer learning concepts
- Evaluating model performance

Module 8: Building ML Pipelines (MLOps)

- Introduction to MLOps
- Creating pipelines in Vertex AI Pipelines
- Automating workflows
- CI/CD for ML models

Module 9: Data Management for GenAI

- Data preparation and preprocessing
- Using BigQuery for AI workloads

- Feature engineering basics
- Data governance

Module 10: Integrating Generative AI Applications

- Building AI-powered apps
- Using REST APIs and SDKs
- Chatbots and content generation apps
- Integration with web and mobile platforms

Module 11: Monitoring and Optimization

- Model monitoring and logging
- Performance tuning
- Cost optimization strategies
- Scaling AI workloads

Module 12: Responsible AI and Security

- Ethical AI principles
- Bias detection and mitigation
- Data privacy and compliance
- Securing AI models and APIs

Module 13: Real-World Use Cases

- Generative AI in business applications
- Case studies (marketing, healthcare, finance)
- Designing enterprise AI solutions

Module 14: Advanced Topics

- Multi-modal AI models
- Retrieval-Augmented Generation (RAG)
- Vector databases and embeddings

- AI agents and automation

Module 15: Capstone Project

- End-to-end Generative AI solution
- Deploying a real-world application
- Performance evaluation and optimization
- Final project presentation

Phone: +91-999-911-1686 **Mail Us:** info@ssdntech.com