



Build Production-Ready Agents on Google Cloud

Course description

In this comprehensive course, you will learn the essential concepts, best practices, and practical techniques for designing, developing, deploying, and managing production-ready AI agents on Google Cloud. You will gain a deep understanding of how to leverage Google Cloud's powerful Generative AI services, foundation models, and MLOps tools to create intelligent agents capable of performing complex tasks, interacting with various systems, and operating reliably at scale. The course emphasizes architectural patterns, data integration, advanced prompt engineering, security, and responsible AI practices, ensuring you can build robust and maintainable agent solutions for real-world business needs.

- Course level: Advanced
- Duration: 2 days

Activities

This course includes presentations, hands-on labs, architectural deep dives, demonstrations, and group exercises.

Course objectives

In this course, you will learn to:

- Design scalable and robust AI agent architectures leveraging Google Cloud services.
- Utilize Google Cloud's foundation models (e.g., Gemini) and Vertex AI for agent development.
- Implement advanced prompt engineering and orchestration techniques (e.g., LangChain) to create intelligent agent behaviors.
- Integrate external data sources using Retrieval Augmented Generation (RAG) on Google Cloud.
- Deploy, monitor, and manage production-grade AI agents using Google Cloud MLOps tools.
- Apply security, governance, and responsible AI principles to agent development.
- Optimize agent performance and cost for enterprise-grade applications.

Intended audience

This course is intended for technical professionals involved in AI/ML development and cloud solutions:

- AI/ML Engineers
- Machine Learning Engineers
- Solutions Architects
- Data Scientists
- Software Developers with an interest in AI
- Cloud Engineers supporting AI initiatives
- Technical Leads overseeing AI projects

Prerequisites

We recommend that attendees of this course have:

- Intermediate proficiency in Python programming.
- Familiarity with Google Cloud Platform (GCP) fundamentals (e.g., IAM, Cloud Storage, basic compute services).
- Basic understanding of machine learning and generative AI concepts.

Course outline

Module 1: Introduction to AI Agents on Google Cloud

- What are AI Agents? Definition, components, and types
- The value proposition of building agents on Google Cloud

- Overview of Google Cloud's Generative AI ecosystem (Vertex AI, Gen AI Studio, Foundation Models)
- Common agent architectural patterns: Orchestration, tools, memory, and reasoning
- Demo: Exploring a simple agent interaction within Vertex AI Generative AI Studio

Module 2: Building Blocks: Foundation Models and Prompt Engineering

- Deep dive into Google's Foundation Models: Gemini, PaLM 2, Codey, Imagen
- Accessing and managing models via Vertex AI SDK and client libraries
- Advanced prompt engineering techniques for agent instructions and persona definition
- Prompt templating, few-shot prompting, and chain-of-thought prompting for complex reasoning
- Hands-on Lab: Interacting with Gemini via Vertex AI SDK and advanced prompt experimentation

Module 3: Agent Orchestration with LangChain on Google Cloud

- Introduction to LangChain and its core components: Chains, Agents, Tools, Memory
- Integrating LangChain with Vertex AI models and custom endpoints
- Designing and implementing custom tools for agent interaction with external APIs and services
- Function calling and structured output with Google Cloud Generative AI models
- Hands-on Lab: Building a tool-using agent with LangChain, Vertex AI, and a custom tool

Module 4: Data Integration and Retrieval Augmented Generation (RAG) for Agents

- The necessity of RAG for knowledge-intensive agents
- Google Cloud services for RAG: Vertex AI Matching Engine (vector search), AlloyDB/Cloud SQL with pgvector
- Embedding models on Vertex AI and optimizing embedding generation
- Designing and implementing RAG pipelines: data ingestion, chunking, indexing, and retrieval
- Hands-on Lab: Implementing a RAG-enabled agent to answer questions from a custom knowledge base

Module 5: Advanced Agent Capabilities: Memory, State, and Human-in-the-Loop

- Managing agent memory: Conversational buffer, summary memory, entity memory, and long-term memory
- Maintaining agent state across multiple interactions and sessions
- Strategies for human-in-the-loop intervention and feedback mechanisms
- Robust error handling, fallback mechanisms, and graceful degradation for agents
- Demo: Implementing and observing different memory strategies and human handoffs

Module 6: Deployment, Monitoring, and MLOps for Agents

- Containerizing AI agents with Docker for deployment
- Deployment options on Google Cloud: Cloud Run, GKE, Vertex AI Endpoints, Cloud Functions
- Monitoring agent performance, usage, and cost with Cloud Monitoring and Logging
- Implementing A/B testing and canary deployments for agent updates
- Establishing CI/CD pipelines for agent development and deployment workflows
- Hands-on Lab: Deploying a production-ready agent to Cloud Run and setting up basic monitoring

Module 7: Security, Governance, and Responsible AI for Agents

- Securing agent endpoints, data access, and model interactions using Google Cloud IAM
- Data privacy, compliance (e.g., GDPR, HIPAA), and data residency considerations
- Principles of Responsible AI for agent development: Fairness, transparency, interpretability, safety
- Mitigating adversarial prompts, prompt injections, and data leakage risks
- Content moderation and safety features with Vertex AI Safety attributes
- Case Study Discussion: Analyzing ethical dilemmas and responsible design choices in agent applications

Module 8: Optimization, Scaling, and Future Trends

- Performance tuning and cost optimization strategies for AI agents

- Scaling agents for high-traffic and concurrent user loads on Google Cloud
- Emerging trends in agentic AI, personal agents, and Google Cloud's roadmap
- Course Wrap-up: Next steps, additional resources, and project planning considerations for enterprise agent initiatives

Your course overview ends here.
But your cloud journey doesn't!

Unlock advanced training and certification pathways with **CloudThat**.



About CloudThat

Established in 2012, CloudThat is an award-winning company and the first in India to offer cloud training and consulting services for individuals and enterprises worldwide. Recently, it won Google Cloud's New Training Partner of the Year Award for 2025, becoming the first company in the world in 2025 to hold awards from all three major cloud giants: AWS, Microsoft, and Google. CloudThat notably won consecutive AWS Training Partner of the Year (APJ) awards in 2023 and 2024 and the Microsoft Training Services Partner of the Year Award in 2024, bringing its total award count to an impressive 12 awards in the last 8 years.

CloudThat specializes in Cloud Migration, Data Platforms, DevOps, IoT, and cutting-edge technologies like Gen AI & AI/ML. As a Microsoft Solutions Partner, AWS Premier Tier Training Partner, Google Cloud Platform Partner, and collaborator with leading organizations like HPE, Oracle, Nvidia and Databricks, CloudThat has trained over 850,000 professionals across 600+ cloud certifications, empowering students and professionals worldwide to advance their skills and careers.

To know more about our cloud certification training, email at sales@cloudthat.com or call us at [+91-8880002200](tel:+91-8880002200) or [+1-8555588830](tel:+1-8555588830) for those from outside India.