



AI Infrastructure Essentials

Course description

Learn the foundational infrastructure components required to build, train, and deploy AI workloads at scale. Explore Google Cloud's AI Hypercomputer architecture, compute accelerators such as GPUs and TPUs, high-performance networking and storage solutions, and deployment models designed to maximize AI performance, scalability, and cost efficiency.

Participants will explore Google Cloud's AI Hypercomputer architecture, learn how to select the right compute accelerators for AI workloads, evaluate networking and storage architectures that maximize training performance, and compare deployment and consumption models for resource optimization. Through discussions, exercises, and practical examples, learners gain the knowledge required to make informed infrastructure decisions for AI initiatives.

- Course level: Intermediate
- Duration: 1 day

Activities

This course includes instructor-led training, architecture discussions, infrastructure design reviews, interactive exercises, scenario-based learning, performance optimization discussions, knowledge checks, and quiz and reflection activities to reinforce learning.

Course objectives

In this course, you will learn to:

- Differentiate between the layers of the AI Hypercomputer.
- Understand the infrastructure requirements of AI workloads.
- Select appropriate GPUs and TPUs for AI use cases.
- Evaluate storage and networking architectures for AI training.
- Understand AI data pipelines and training workflows.
- Optimize training performance through improved goodput.
- Compare deployment and consumption models for AI infrastructure.
- Make informed infrastructure decisions for enterprise AI projects.
- Align AI infrastructure investments with business requirements.

Intended audience

This course is intended for:

- IT Decision Makers
- Infrastructure Architects
- Cloud Architects
- Enterprise Architects
- Technology Leaders
- Platform Engineers
- Technical Managers
- AI Infrastructure Specialists
- Cloud Operations Teams

Prerequisites

We recommend that attendees of this course have:

- Familiarity with cloud computing concepts.
- Understanding of general data center infrastructure.
- Basic knowledge of enterprise IT architecture is beneficial.
- Interest in AI infrastructure and cloud technologies.

Products covered

- AI Hypercomputer
- Google Cloud GPUs
- Google Cloud TPUs
- Google Cloud Networking Services
- Google Cloud Storage Services

Course outline

Module 01 - Foundations of AI Infrastructure

- Definition of AI Infrastructure
- Evolution of Computing Demands
- Increasing Requirements for AI Workloads
- Modern AI Infrastructure Challenges

Module 02 - Google Cloud's AI Hypercomputer

- Introduction to AI Hypercomputer
- AI Hypercomputer Architecture
- The Three Layers of the AI Hypercomputer
- Infrastructure Optimization Principles

Module 03 - Compute Accelerators: GPUs and TPUs

- Graphics Processing Units (GPUs)
- GPU Architecture
- Google Cloud GPU Family
- GPU Selection Criteria
- Tensor Processing Units (TPUs)
- TPU Architecture
- Google Cloud TPU Family
- Best Practices and Considerations

Module 04 - The AI Data Pipeline: Network and Storage

- Maximizing Training Goodput
- Networking for Data Ingestion

- Networking for AI Training
- Storage for Data Preparation
- Storage for AI Training
- Infrastructure for AI Inference

Module 05 - Orchestration and Consumption

- AI Infrastructure Deployment Options
- Flexible Consumption Models
- Resource Allocation Strategies
- Infrastructure Utilization Optimization

Module 06 - Course Summary and Quiz

- Course Summary
- Review of Key Concepts
- Q&A Session
- Knowledge Assessment Quiz

Certification details

Participants receive a course completion certificate upon successfully attending the training program.

FAQs

Who should attend this course?

This course is designed for IT decision-makers, infrastructure architects, cloud architects, and technology leaders responsible for AI infrastructure planning and deployment.

What is the AI Hypercomputer?

AI Hypercomputer is Google's integrated AI infrastructure architecture designed to optimize compute, networking, and software layers for large-scale AI training and inference workloads.

Does the course cover GPUs and TPUs?

Yes. The course covers GPU and TPU architectures, Google Cloud accelerator families, selection criteria, and best practices.

What is training goodput?

Training goodput refers to the effective performance achieved during AI training after accounting for bottlenecks, inefficiencies, and infrastructure limitations.

Does the course cover AI networking and storage?

Yes. Networking, storage, data ingestion, training pipelines, and inference architectures are key topics covered in the course.

Is this a hands-on course?

The course focuses on infrastructure concepts, architecture reviews, discussions, and practical examples rather than hands-on labs.

How long is the course?

The course is delivered in a 3-hour instructor-led format.

What business benefits can organizations expect?

Organizations can improve AI infrastructure planning, optimize infrastructure investments, reduce operational costs, improve training performance, accelerate AI adoption, and build scalable foundations for enterprise AI initiatives.