

AI Infrastructure Essentials

This course provides a foundational overview of the hardware, software, and networking components required to develop and manage AI models at scale. It explores Google Cloud's AI Hypercomputer architecture, compares compute accelerators like GPUs and TPUs, and examines the critical data pipelines and storage solutions necessary to maximize training performance.

**DURATION**

2-3 hours

**LEVEL**

200 - Intermediate

**FORMAT**

Instructor-led

**SCALABLE**

1-50, scaled to 1000

What you'll learn

- Differentiate between the layers of the AI Hypercomputer.
- Select appropriate accelerators for the most cost-effective AI workloads.
- Evaluate storage and networking solutions to maximize training goodput.
- Compare various deployment and consumption models for resource optimization.

Overview	6 modules
Who this course is for	IT decision-makers and infrastructure architects looking to understand the technical requirements and the AI Hypercomputer's offerings for enterprise-grade AI deployment.
Products	- AI Hypercomputer - Storage Insights - Cloud TPU v5 Lite Pod
Prerequisites	Familiarity with cloud computing concepts and general data center infrastructure.

Module 01 Foundations of AI infrastructure

Topics	- Definition of AI infrastructure - The evolution of computing demands - The need for new computing power
--------	---

Module 02 Google Cloud's AI Hypercomputer

Topics	- The AI Hypercomputer - The 3 layers of the AI Hypercomputer: Overview
Objectives	Differentiate between the layers of the AI Hypercomputer.

Module 03

Compute accelerators: GPUs and TPUs

Topics	• Graphics Processing Units • GPU architecture • Google Cloud GPU family • Selecting GPUs • Tensor Processing Units • TPU architecture • Google Cloud TPU family • Best practices and considerations
Objectives	• item
Activities	• 1x exercise/discussion

Module 04

The AI data pipeline: Network and storage

Topics	• Maximizing goodput • Networking for data ingestion and training • Storage for data preparation and training • Architecture for inference
Objectives	• Evaluate storage and networking solutions to maximize training goodput.
Activities	• 1x discussion

Module 05

Orchestration and consumption

Topics	• Deployment options • Flexible consumption
Objectives	• Compare various deployment and consumption models for resource optimization.

Module
06

Course summary and quiz

Topics	• Course summary • Q&A • Quiz
Objectives	• Differentiate between the layers of the AI Hypercomputer. • Select appropriate accelerators for the most cost-effective AI workloads. • Evaluate storage and networking solutions to maximize training goodput. • Compare various deployment and consumption models for resource optimization.
Activities	• 1x quiz with 4 MCQs