Google Skills

Exploring your orchestration options

While we've outlined two of the primary paths, your approach will vary based on model complexity, performance needs, your team's preference, and operational capacity. For example, you could start with fully managed solutions and progressively adopt more customized infrastructure as your requirements evolve, or you could use several options in tandem for jobs with different requirements. Let's explore the available options.

Option 1

Use pre-trained AI APIs and pre-built models for rapid deployment

This approach is ideal for developers of any skill level, including those new to AI, who want to integrate powerful AI capabilities quickly. It requires making simple API calls without needing to manage any models or infrastructure.

Vertex AI Model Garden

Use Google's Gemini models and a selection of open-source models with a simple API endpoint. It handles the complexities of hosting and scaling so you can focus on your application and get powerful results for generative AI tasks.

Option 2

Perform inference directly in your data warehouse using SQL

Those working with SQL can now get predictions from AI models right where your data already lives. This simplifies your workflow by eliminating the need to move data to a separate platform.

BigQuery ML

BigQuery allows you to run machine learning models directly on your data with simple SQL commands, eliminating the need to move data and reducing complexity and latency. It's a highly efficient method for batch processing tasks like customer segmentation or demand forecasting, especially when your data is already stored in BigQuery.

Option 3

Build a custom serving infrastructure for maximum control

This option gives developers and MLOps granular control and flexibility to deploy, manage, and scale custom containerized inference services—often with specialized hardware across cloud or hybrid environments.

Google Kubernetes Engine (GKE)

GKE is the selection made by most. It's a Kubernetes-based service that provides **more** control over hardware, including CPUs, GPUs, and TPUs, while automating lower-level management tasks like node upgrades, scaling, node repair, and deployment.

Cluster Director (CD)

CD simplifies deploying, scaling, and managing AI workloads while automating lower-level management tasks. CD is ideal if you don't want to use GKE, but would still like to be able to set up optimized multi-host AI clusters quickly (for example, for Slurm).

Google Compute Engine (GCE)

GCE provides **complete** control over your hardware, allowing you to design your stack from the VM up. This approach requires the most work, but is ideal for those with highly customized infrastructure requirements.

Option 4

Deploy custom models on a fully-managed service

This option is for developers who already have a custom model built. Deploying to our managed services lets you skip complex server setup and orchestration, so you can focus on your model without worrying about infrastructure.

Vertex AI Prediction

Vertex AI Prediction is a managed service that deploys machine learning models as scalable endpoints, using hardware accelerators like GPUs for fast processing of both real-time and large-batch data.

Cloud Run

Deploying containerized models with auto-scaling to zero and pay-per-request pricing is ideal for highly variable intermittent workloads or simple web services.

While there are several available options, most of the content and examples in this course will use GKE. Not only do the vast majority of inference users prefer Kubernetes, but GKE is also an optimal environment to run it.