Google Skills

Best practices, tutorials and ideas from Google Developer Experts (GDEs)

GDEs are a community of 1000+ experts (customers, not Googlers) who have achieved the highest level of certification available. To become a GDE they must be nominated by a Googler and demonstrate deep technical expertise in Google Cloud. We've collated some relevant GDE content below:

The Enterprise Guide to Scaling AI Agents on Google Cloud Platform:

A comprehensive guide explores the strategic and technical considerations for implementing AI agents at scale, providing a roadmap for organizations ready to embrace the next generation of intelligent automation.

GKE Agent Workshop:

This github project demonstrates an AI-powered agent capable of managing Google Kubernetes Engine (GKE) clusters and Docker images. The agent leverages the Google Agent Development Kit (ADK) and custom tools to interact with Google Cloud services.

Orchestrating the Container Symphony - From Pods to Production on GKE:

Explores core concepts about Kubernetes like Pods, Deployments, Services, and Jobs, understanding how they orchestrate containerised applications.

To implement the GKE reference architecture

1. Deploy from scratch

A complete, production-ready reference implementation is available as Terraform code. Find instructions and deploy it from the official GitHub repository via the GKE Inference Reference Implementation Guide.

2. Deploy using one of these more opinionated use cases

Scalable and Distributed LLM Inference on GKE with vLLM

- **Efficient deployment with vLLM:** Leveraging single GPU, single-node multi-GPU strategies
- **Optimizing performance:** Accelerating container image pulls and model weight loading
- **Inference modes:** Understanding batch and real-time inference and their respective use cases
- **Production monitoring:** Using Prometheus and custom metrics for observability
- **Scaling strategies:** Dynamically scaling your deployment with Horizontal Pod Autoscaler (HPA)

Create a model fine-tuning pipeline

- Data set preparation
- Fine-tuning of the instruction tuned model
- Hyperparameter tuning to generate multiple model iterations
- Model validation, evaluation, and identification of the optimal model

Build a Retrieval Augmented Generation (RAG) pipeline to combine a pre-trained large language model with semantic search to retrieve relevant information from the product catalog and provide answers based on the relevant information.

Additional reading materials:

- Recommended configurations
- How to obtain capacity
- Choose a deployment strategy
- About AI/ML model inference on GKE
- Choose a load balancing strategy for AI/ML model inference on GKE
- Manage cluster lifecycle changes to minimize disruption
- Optimize your usage of GKE with insights and recommendations
- Observability for GKE
- AI Hypercomputer GitHub