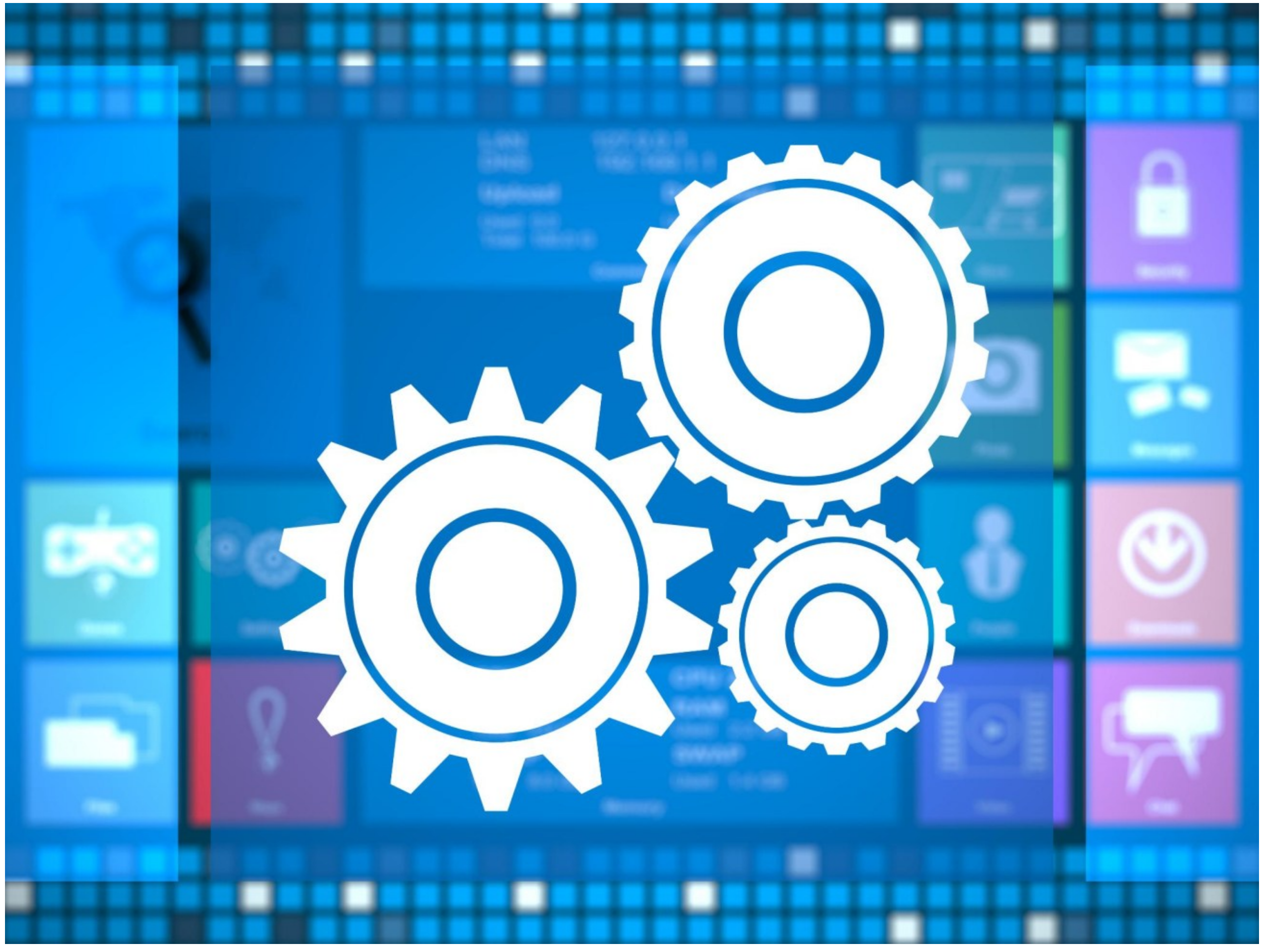
Day 8: Deployment Options in Google ADK

Exploring various deployment methods for ADK applications



DEPLOYMENT OPTIONS OVERVIEW

Running Locally vs Cloud Deployment

Local Deployment Benefits

Running AI locally enables quick testing without internet and supports rapid development cycles.

Local Deployment Limitations

Local deployment is limited by machine resources, unsuitable for handling large-scale or production workloads.

Cloud Deployment Advantages

Cloud deployment offers scalability, reliability, managed infrastructure, and global accessibility for production environments.

Choosing Deployment Approach

Selection depends on project stage, expected traffic, and resource availability for optimal performance.

DEPLOY ON VERTEX AI AGENT ENGINE

Vertex AI Agent Engine Deployment

Managed Conversational AI Service

Vertex AI Agent Engine offers efficient deployment of conversational AI with seamless cloud integration and scalability.

Multi-turn Conversation Support

Supports multi-turn dialogue and context retention for sophisticated and natural AI interactions.

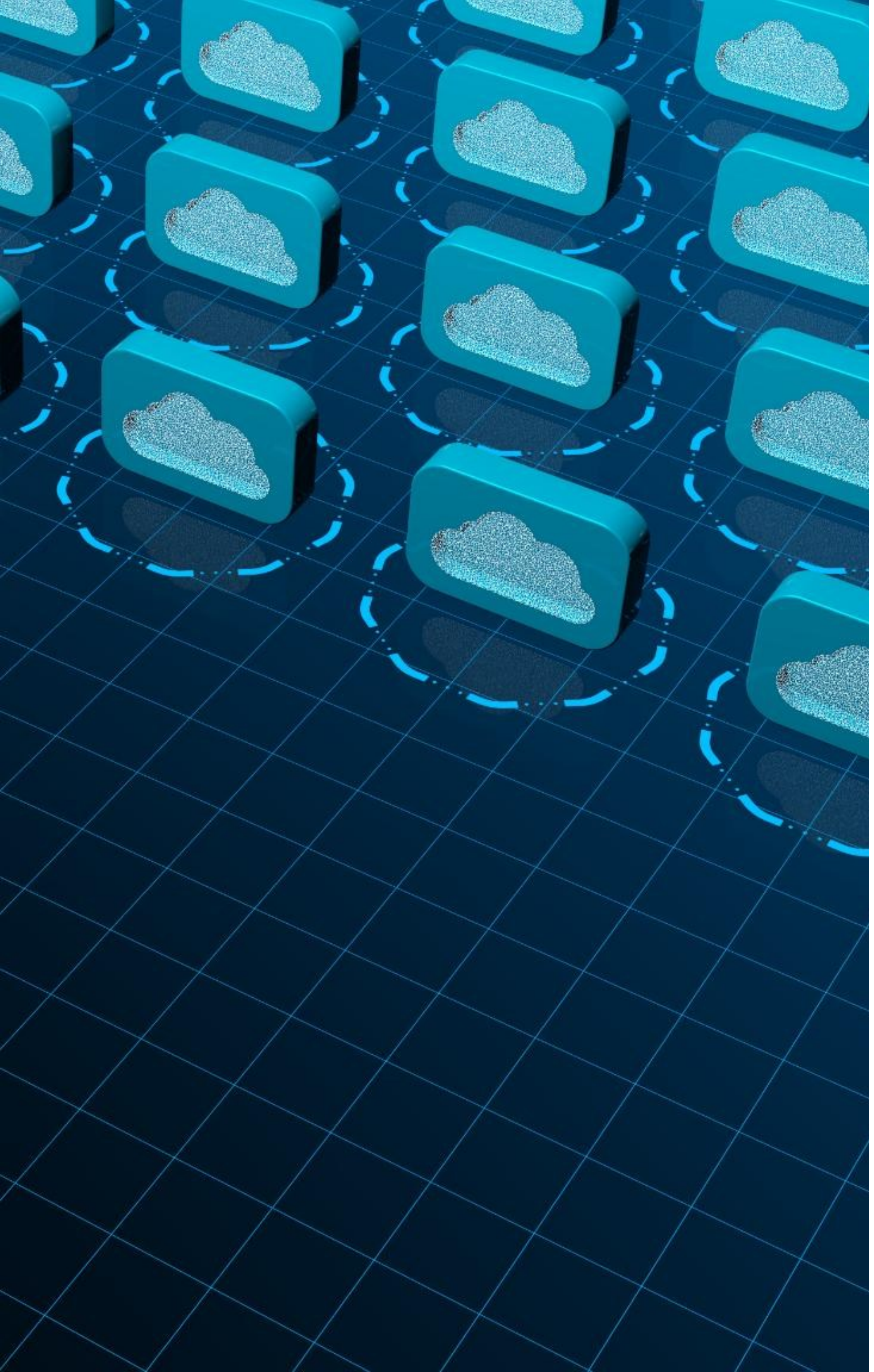
Secure and Compliant Deployment

Built-in security and compliance features ensure enterprise-grade safe deployments using IAM and access controls.

Deployment Workflow

Deployment includes configuring agents, uploading models, setting endpoints, and testing through simulators and dashboards.

CONTAINERIZATION OPTIONS



Cloud Run and GKE for Containerized Deployment

Containerization Benefits

Containerization ensures portability, consistency, and simplifies scaling of AI agents across different environments.

Cloud Run Platform

Cloud Run is a fully managed, serverless platform that scales containers automatically based on traffic, ideal for lightweight deployments.

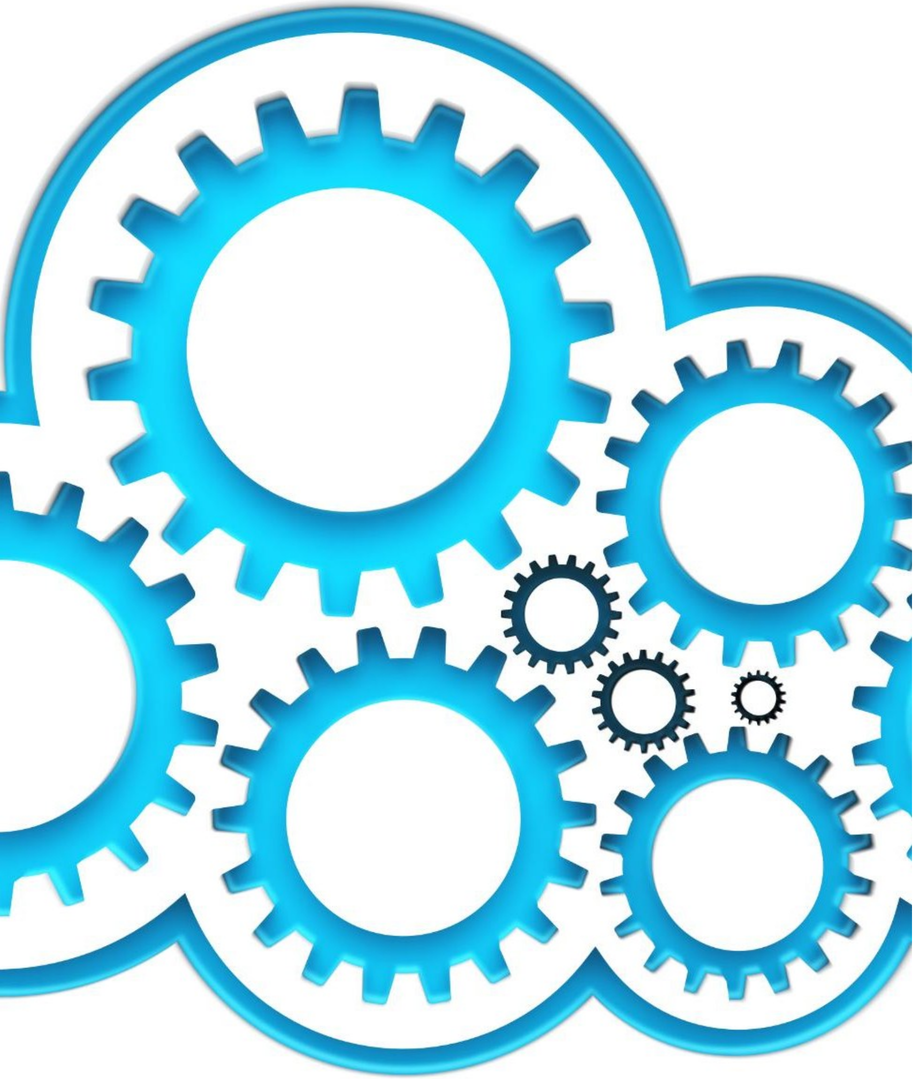
Google Kubernetes Engine (GKE)

GKE offers full control over container orchestration using Kubernetes, suitable for complex workloads requiring custom configurations.

Containerization Workflow

Workflow includes building Docker images, pushing to container registries, then deploying on Cloud Run or GKE.

BEST PRACTICES



Deployment Best Practices

Automation with CI/CD

Use Continuous Integration and Continuous Deployment pipelines to automate builds and testing, reducing errors and speeding releases.

Performance Monitoring

Monitor metrics like latency, throughput, and error rates to ensure optimal system performance and user experience.

Security and Access Control

Apply strict IAM roles and security policies to protect endpoints and prevent unauthorized access.

Cost Efficiency and Testing

Scale resources dynamically to optimize costs and conduct thorough testing in staging before production deployment.