

Q1. What is Vertex AI and what is its primary purpose in its cloud ecosystem?

A: Vertex AI is a managed cloud service that handles a specific infrastructure concern (compute, storage, messaging, AI). Using it shifts operational burden to the cloud provider while you focus on application logic.

Q2. How does IAM work with Vertex AI?

A: Access to Vertex AI is controlled via IAM roles and policies. Attach the minimum required permissions to a service account or role. Avoid long-lived access keys; use instance roles or Workload Identity Federation instead.

Q3. How do you monitor Vertex AI in production?

A: Vertex AI emits metrics to the cloud provider's monitoring service (CloudWatch, Cloud Monitoring, Azure Monitor). Set alerts on key metrics (error rate, latency, capacity). Use distributed tracing to correlate across services.

Q4. How do you scale Vertex AI to handle variable load?

A: Vertex AI supports auto-scaling policies based on CPU, queue depth, or custom metrics. Set minimum and maximum capacity. Horizontal scaling (more instances) is generally preferred over vertical for cloud workloads.

Q5. How do you secure Vertex AI in production?

A: Place Vertex AI in a private subnet with no public endpoint where possible. Use VPC security groups or firewall rules to allow only required traffic. Encrypt data at rest and in transit. Rotate credentials regularly.

Q6. How do you manage configuration and secrets for Vertex AI?

A: Store sensitive values in a managed secret store (AWS Secrets Manager, GCP Secret Manager, Azure Key Vault). Inject secrets as environment variables at runtime, never bake them into container images or source code.

Q7. What are the main cost drivers for Vertex AI and how do you optimize them?

A: Cost in Vertex AI typically comes from compute hours, data transfer, storage, and request counts. Right-size instances, use reserved capacity for steady-state workloads, and enable lifecycle policies to expire old data.

Q8. How do you implement high availability with Vertex AI?

A: Deploy Vertex AI across multiple availability zones. Use managed replicas or active-active configurations. Implement health checks and automatic failover. Test resilience regularly with chaos engineering or failover drills.

Q9. How does Vertex AI integrate with a CI/CD pipeline?

A: CI/CD pipelines use the cloud CLI or SDK to deploy updates to Vertex AI after tests pass. Infrastructure-as-code (Terraform, CDK) defines Vertex AI configuration as version-controlled code deployed via the pipeline.

Q10. What are common pitfalls when first adopting Vertex AI?

A: Common mistakes include misconfigured IAM policies (too permissive), no cost alerting, deploying only in one availability zone, and missing backups. Read the service SLA carefully to understand what the cloud provider guarantees.