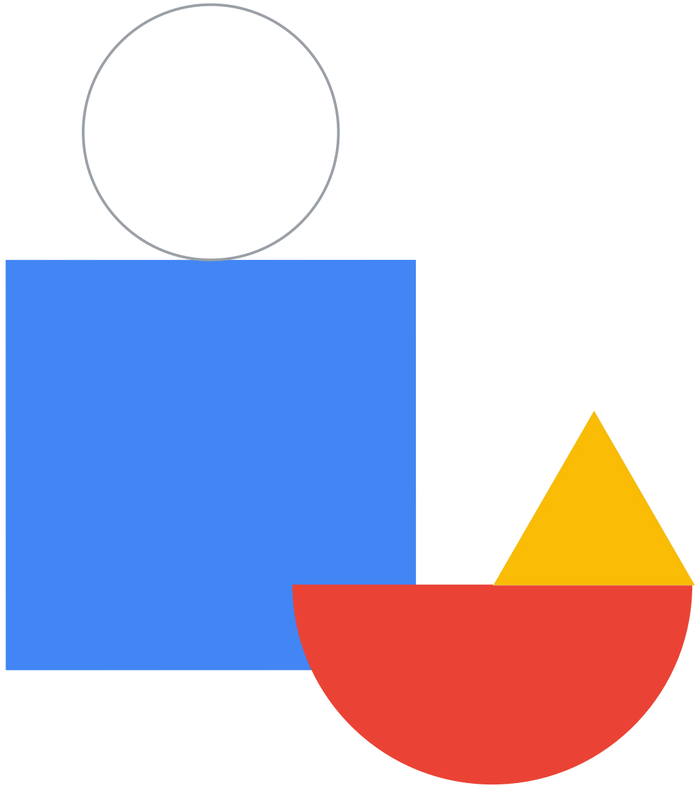
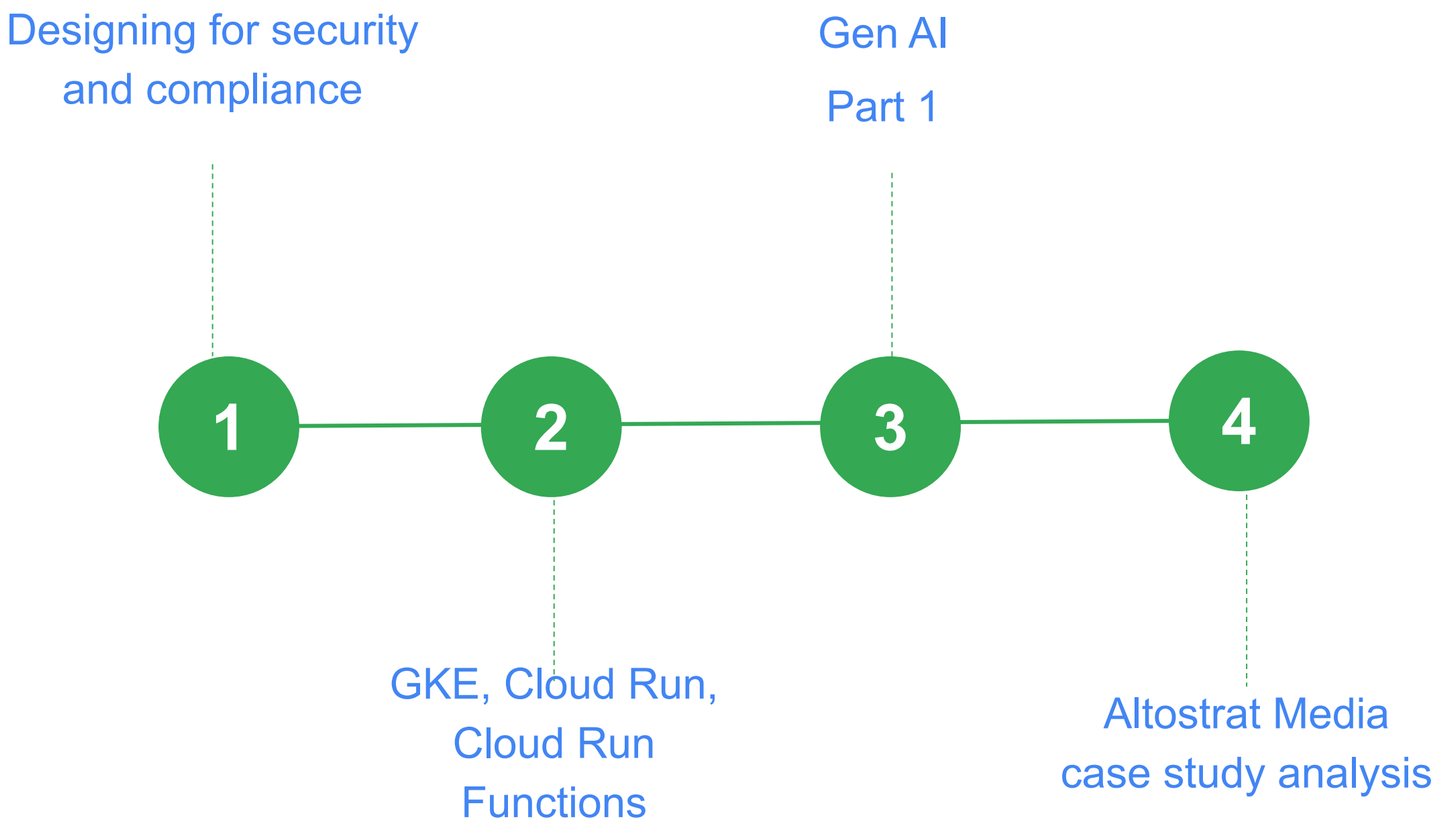
Google Cloud



Preparing for Your Professional Cloud Architect Journey

Module 3: Designing for Security and Compliance

Week 4 topics

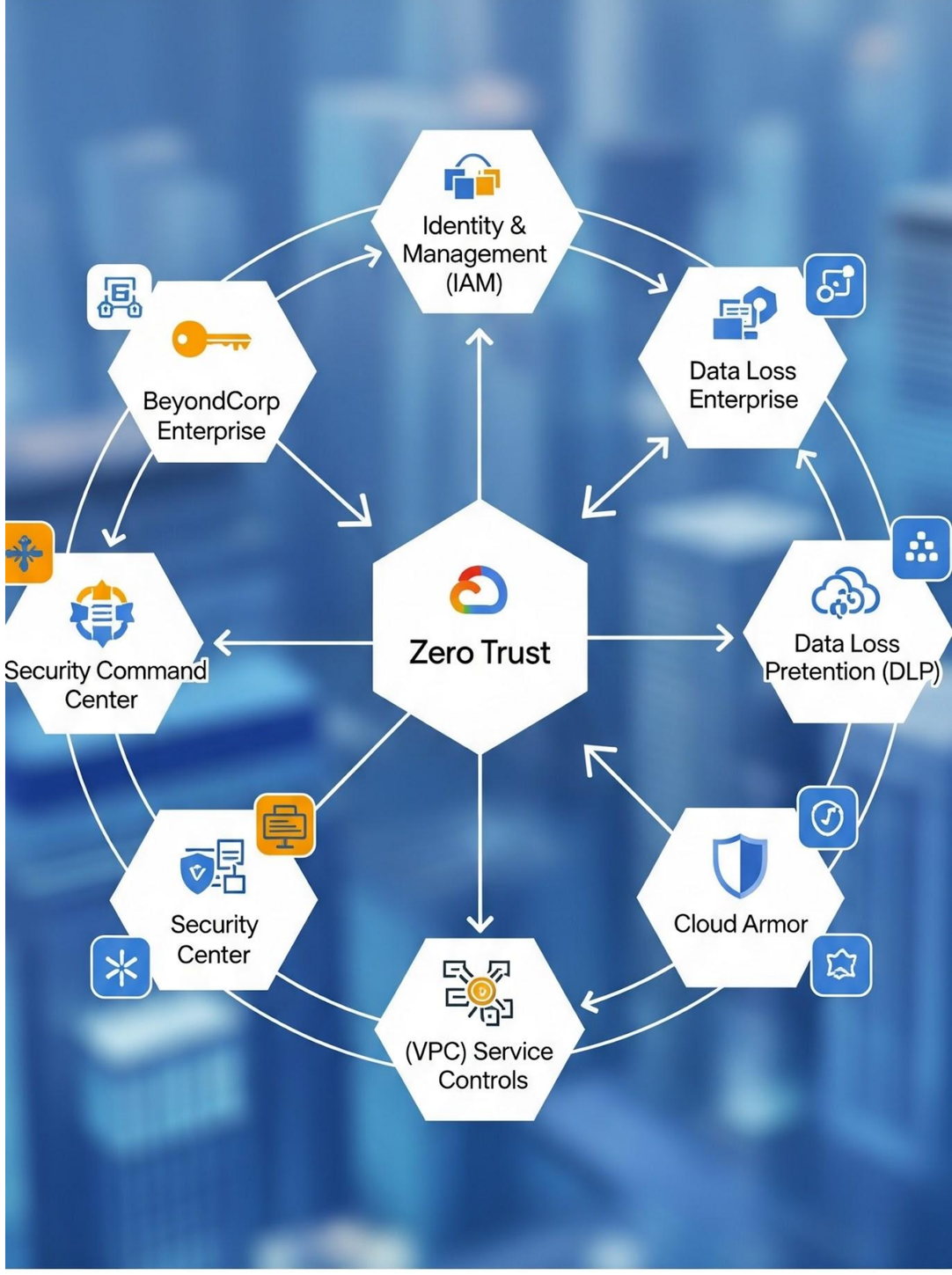


Designing for security and compliance

Google’s approach to Zero Trust

…. and other security measures

- Intro to Zero Trust
  - Zero Trust at Google
- Security by design
- Shared responsibility and shared fate
- Shifting security left
- Secure and responsible AI
- How to use AI for security



Compliance in GCP - 1/2

- **ISO 27001**
  - Requirements for an information security management system (ISMS), specifies a set of best practices
  - ONLY GUIDANCE, lays out allow Google to ensure a comprehensive and continually improving model for security management.
- **SOC 2**
  - The purpose of this report is to evaluate an organization's information systems relevant to security, availability, processing integrity, confidentiality, and privacy.
  - Relevant are different services: VPC Service Controls, DLP, Cloud Security Command Center, Cloud Armor etc
- **PCI DSS**
  - Appropriate practices that merchants and service providers should follow to protect cardholder data.
  - Relevant are MANY GCP services: networking, logging, encryption etc
- **FIPS 140-2**
  - A security standard that sets forth requirements for cryptographic modules, including hardware, software, and/or firmware, for U.S. federal agencies.
  - Google Cloud Platform uses a FIPS 140-2 validated encryption module called BoringCrypto (certificate 3318) in our production environment. This means that both data in transit to the customer and between data centers, and data at rest are encrypted using FIPS 140-2 validated encryption.

Compliance in GCP - 2/2

- HIPAA
  - Healthcare-related.
  - Complying with HIPAA is a shared responsibility between the customer and Google.
  - Google Cloud Platform supports HIPAA compliance (within the scope of a Business Associate Agreement) but ultimately customers are responsible for evaluating their own HIPAA compliance.
- FedRAMP
  - Government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.
  - Risk impact levels (Low, Moderate, or High)
  - Google is one of the first hyperscale commercial cloud providers to achieve FedRAMP High on a commercial public cloud offering, and is one of the largest providers of FedRAMP services available on the market today.
  - NO SEPARATE 'GOVERNMENT' REGIONS EXIST IN GCP.
- GDPR
  - PII data protection in Europe.
  - Our customers own their data and we believe they should have the strongest levels of control over data stored in the cloud. Our public cloud provides customers with world-class levels of visibility and control over their data through our services.
  - Storing data in Europe, optionally manage encryption keys and store them outside of GCP, External Key Manager etc.

How do you ensure compliance?

By implementing “security-relevant” options!

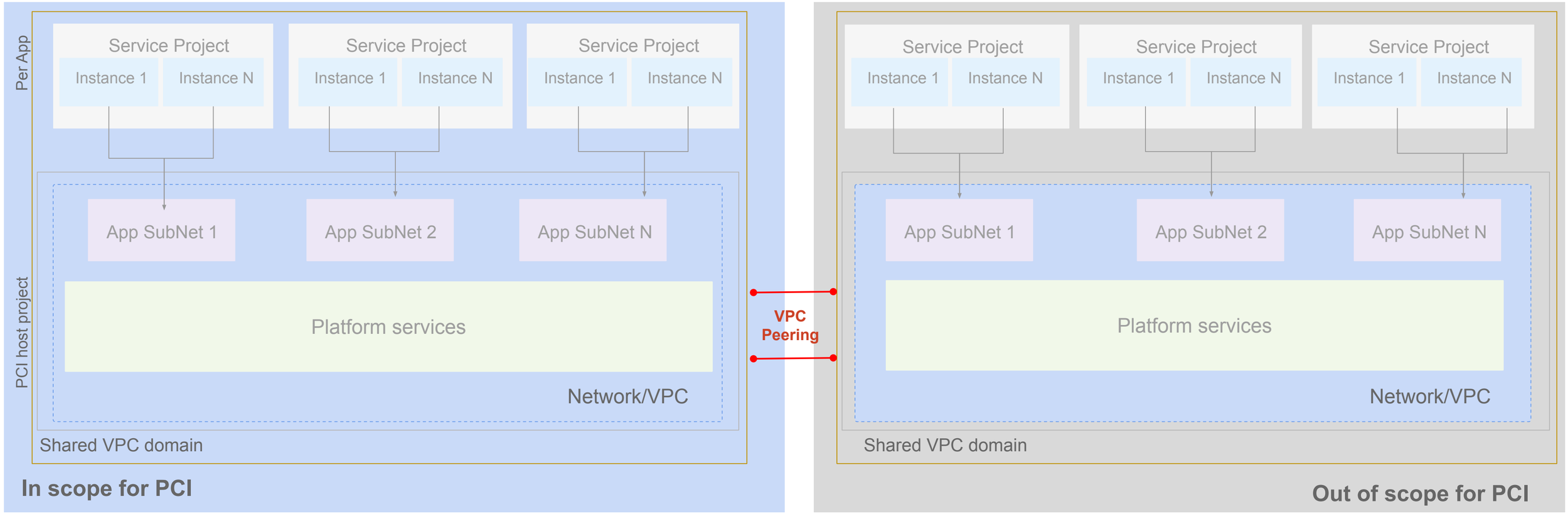
Google Security Overview	Shielded VMs	Identity and Access Management
Access Transparency	Confidential Computing	IAM Conditions
GCP Compliance offerings	Shared VPC	Identity-Aware Proxy
Binary Authorization	VPC Service Controls	Resource Manager
Data Loss Prevention	Cloud Armor	Private Service Connect
Key Management Service	DNSSEC	Private Google Access
Organization Policy Service	Cloud VPN	Serverless VPC Access
Anthos Service Mesh	VPC Flow Logs	Web Security Scanner
Cloud Asset Inventory	Firewall Insights	Cloud Audit Logs
OS Login	Packet Mirroring	Centralized Telemetry

and more…

Mapping PCI-DSS requirements to GCP

Requirement 1

Install and maintain a firewall configuration to protect cardholder data



Architecture - Using Shared VPC, host, and service projects to reduce scope of PCI environment through segmentation of networks. VPC network peering makes services available across VPC networks in private RFC 1918 space using Firewall access control lists.

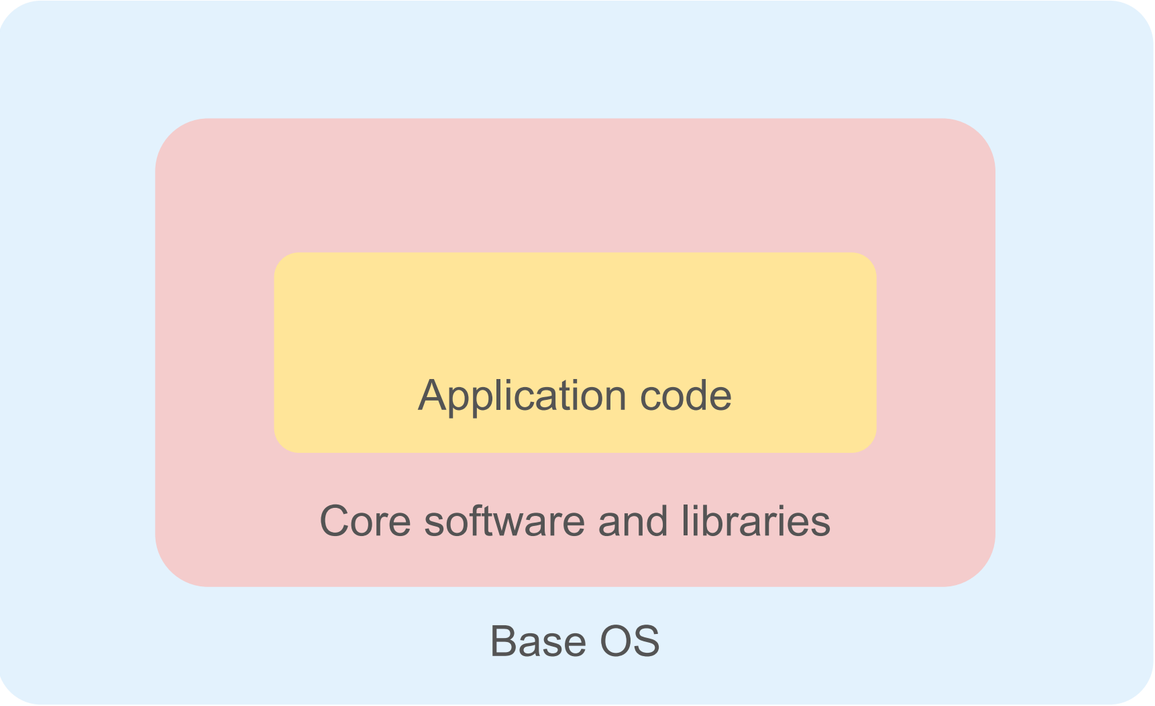
Requirement 2 | Do not use vendor-supplied defaults

Image baking

Base image - OS or hardened image from CIS with unnecessary packages removed

Core - packages and libraries needed for all instances (security, monitoring, language specific packages)

Application - application code



Requirement 3 | Protect stored cardholder data

Enjoy world class encryption without further need for configurations
By default

Keep keys in the cloud, for direct use by cloud services
Generally available

Keep keys on-premises, and use them to encrypt your cloud services
Available for Cloud Storage and Compute Engine

More Simple

Encryption by default (only in GCP)

Cloud key management service

Customer-supplied encryption keys

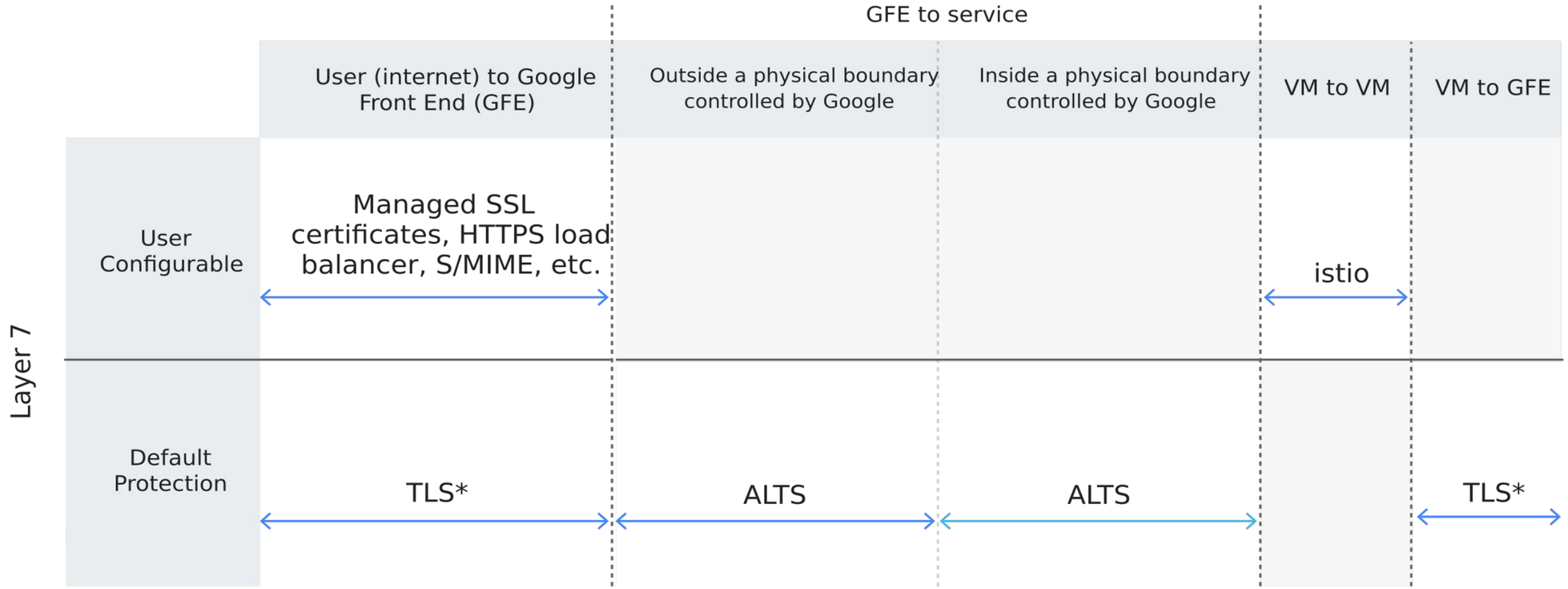
More control

Requirement 3 | Protect stored cardholder data (cont.)

Data Loss Prevention API can be used to sanitize PCI data

Requirement 4

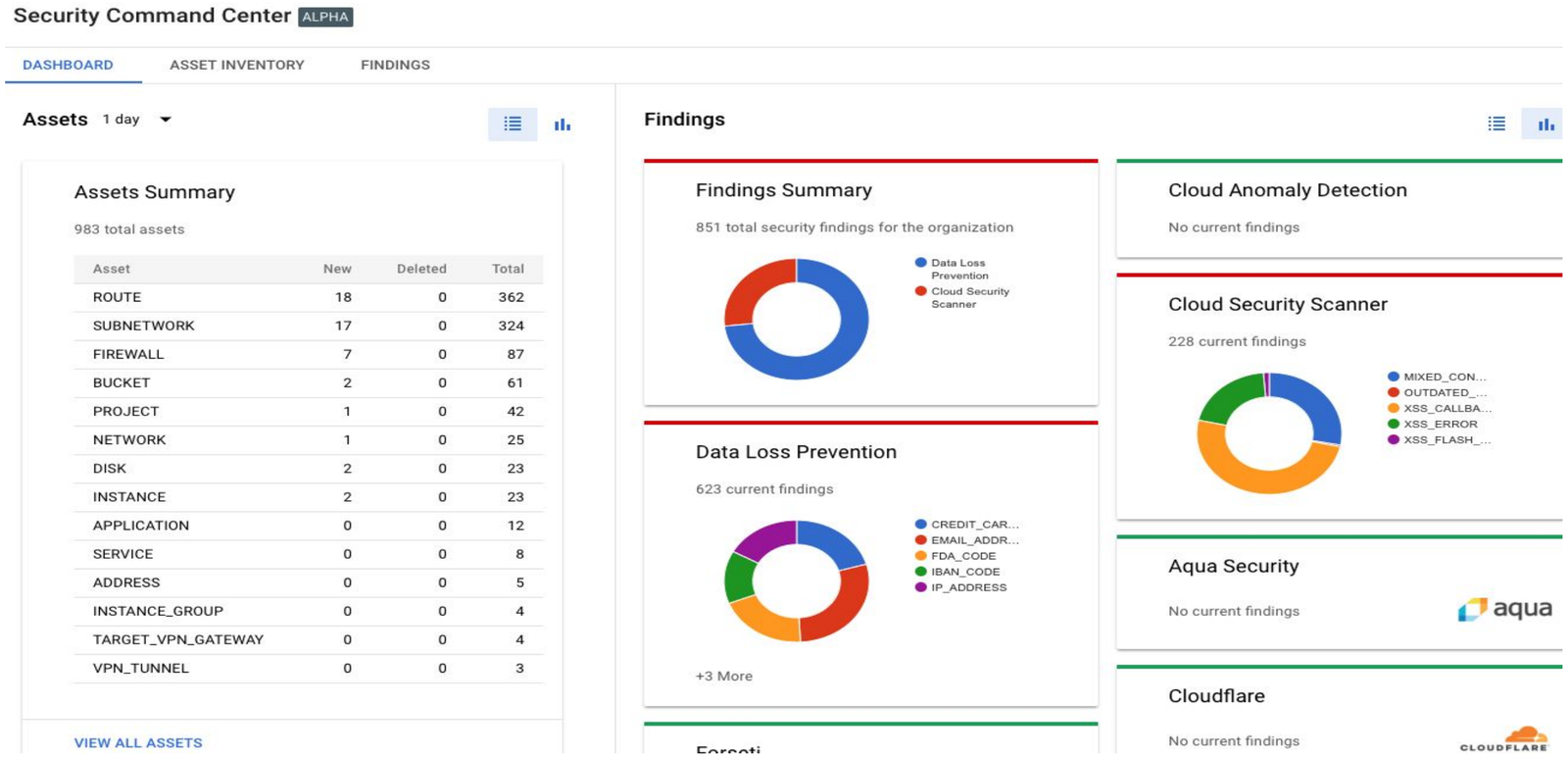
Encrypt transmission of cardholder data across open, public networks



* TLS is by default for Google Cloud services. For a customer application hosted on Google Cloud, this is something that needs to be configured by the customer.

Requirement 5

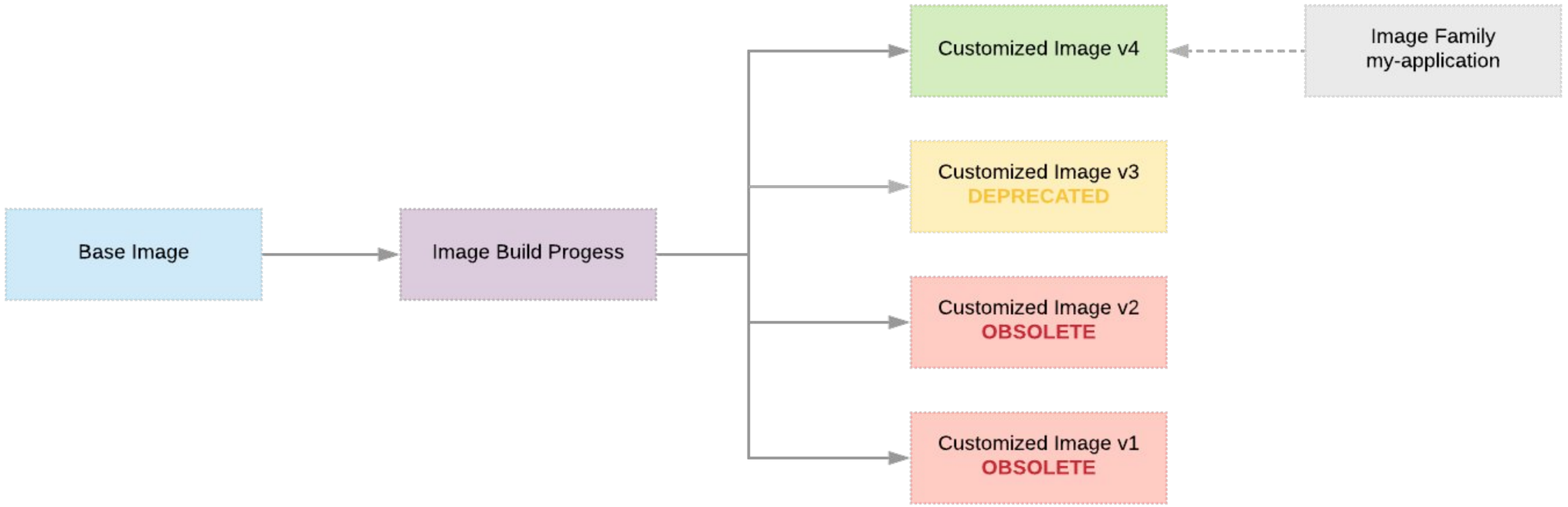
Protect all systems against malware and regularly update anti-virus software or programs



Cloud Security Command Center can help gather security information, identify threats, and take action.

Requirement 6

Develop and maintain secure systems and applications



Deprecate old images so they are not used inadvertently.

[OS Image families best practices]

Requirement 7

Restrict access to cardholder data by business need to know

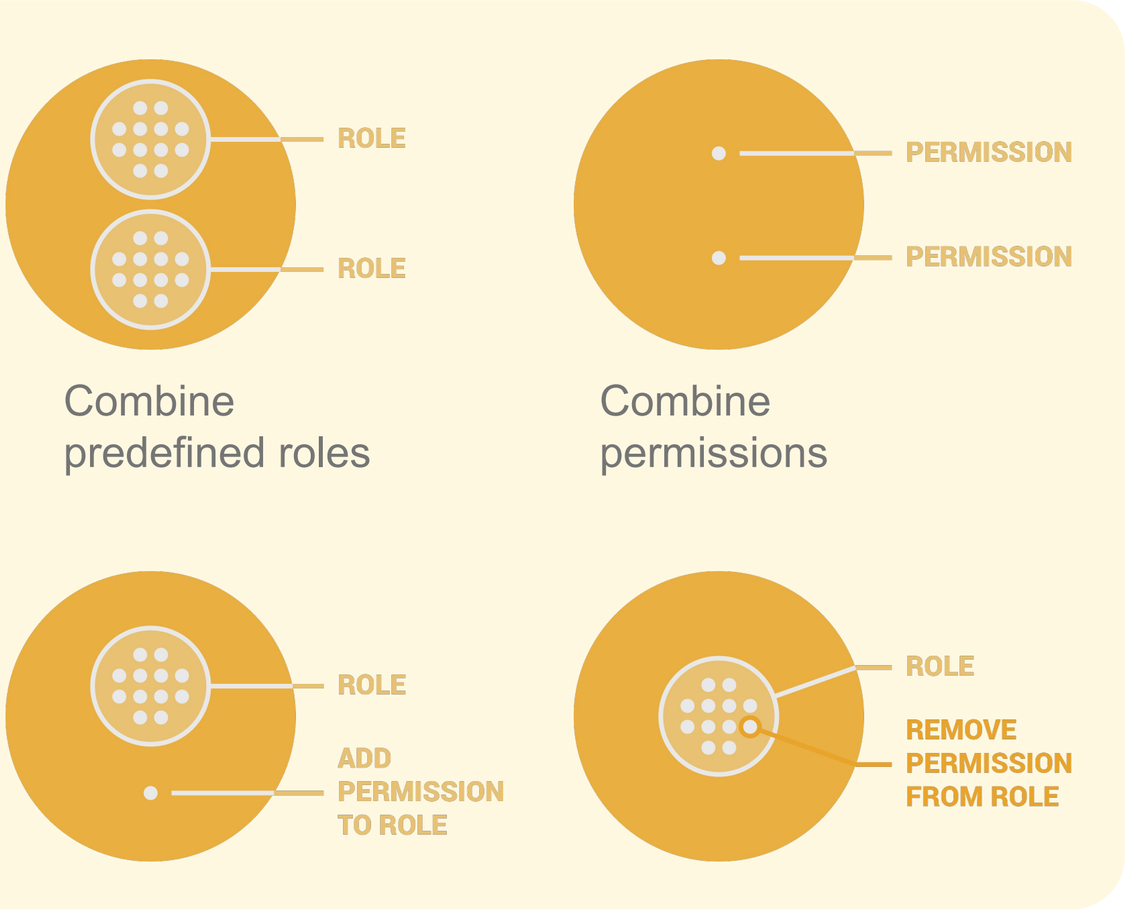
Once access needs for each job function are defined, **custom roles** can be created provide granular control over the exact permissions to access system components and data resources

- Create groups based on job functions, and assign custom roles to those groups.
- Job function groups can be nested in job classification groups.
- Custom roles can be defined at the organizational level

Review available permissions and their purpose through the API Explorer (search for product)

Services > App Engine Admin API v1

appengine.apps.authorizedCertificates.create	Uploads the specified SSL certificate.
appengine.apps.authorizedCertificates.delete	Deletes the specified SSL certificate.
appengine.apps.authorizedCertificates.get	Gets the specified SSL certificate.
appengine.apps.authorizedCertificates.list	Lists all SSL certificates the user is authorized to administer.



Requirement 8

Track and monitor all access to network resources and cardholder data

Admin console logs

- Admin console audits
- User audits
- Separate API and UI
- Export to BigQuery

Cloud audit logs

- Admin activity logs (always enabled)
- Data access logs (disabled by default)

Stackdriver logging agent

- FluentD agent
- Common third-party applications
- System software

Network logs

- VPC flow
- CDN (Alpha)
- HTTP(S) load balancing (Alpha)
- Firewall rules logging

Google Kubernetes Engine (GKE)

***Exam Tip**: I can't stress enough how important it is to understand Kubernetes concepts. Commit at least few hours for learning GKE - especially if you're not familiar with this technology. Slides below will give you a high level overview, but you should be much more familiar with this topic to feel comfortable during the exam.*

GKE: Autopilot Mode

GKE manages underlying infrastructure of the cluster, including the nodes

High availability

Regional cluster; Regular Release Channel; Auto-Update; Auto-Repair; Surge Upgrade;

Network

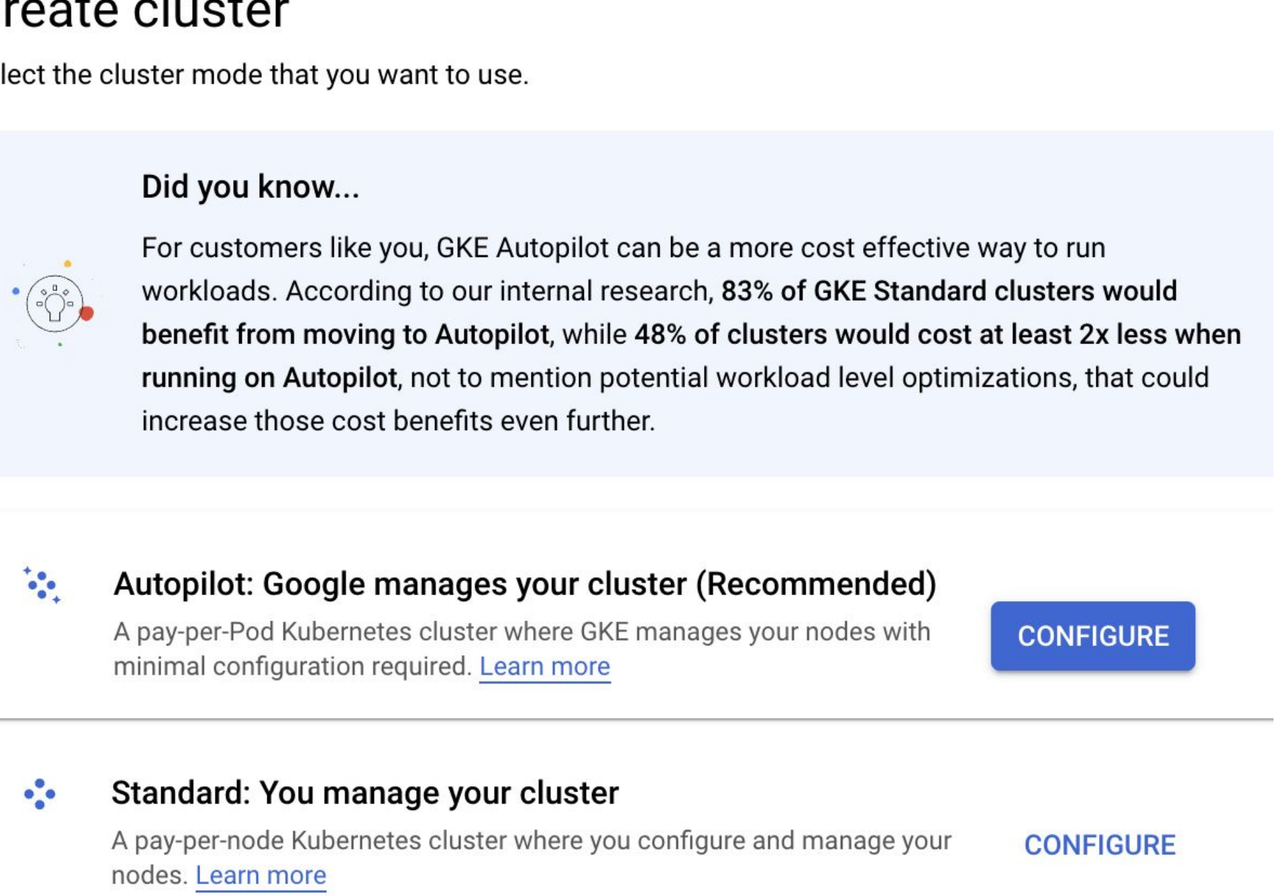
VPC Native (alias IP); IP-friendly (limit cluster size/ pods per node); full network flexibility

Highly Scalable

Node Auto Provision; Horizontal Pod Autoscaler; Vertical Pod Autoscaler

Secured by default

Workload Identity; Shielded Nodes; Secure-boot-disk; COS and Containerd, block known unsecure features.



Pod Disruption Budget, Readiness and Liveness Probes

A **PDB (Pod Disruption Budget)** limits the number of pods of a replicated application that can be taken down **simultaneously** from **voluntary** disruptions.

An Application Owner can create a **PodDisruptionBudget** object (PDB) for each application.

Readiness probes: designed to know when your app is ready to serve traffic.

Liveness probes: designed to let Kubernetes know if your app is alive or dead.

```
kind: PodDisruptionBudget
metadata:
 name: km-pdb
spec:
 minAvailable: 2
 selector:
   matchLabels:
     app: kobimysql
```

```
maxUnavailable: 1
```

Exam Tip:

- *See how to ensure stateful workloads are disruption-ready*
- *Great explanation of Readiness and Liveness probes here.*

Diagnostic Question Discussion

Your company has an application running as a Deployment in a Google Kubernetes Engine (GKE) cluster. When releasing new versions of the application via a rolling deployment, the team has been causing outages. The root cause of the outages is misconfigurations with parameters that are only used in production. You want to put preventive measures for this in the platform to prevent outages.

What should you do?

A. Configure liveness and readiness probes in the Pod specification.

B. Configure health checks on the managed instance group.

C. Create a Scheduled Task to check whether the application is available.

D. Configure an uptime alert in Cloud Monitoring.

Diagnostic Question Discussion

Your company has an application running as a Deployment in a Google Kubernetes Engine (GKE) cluster. When releasing new versions of the application via a rolling deployment, the team has been causing outages. The root cause of the outages is misconfigurations with parameters that are only used in production. You want to put preventive measures for this in the platform to prevent outages.

What should you do?

A. **Configure liveness and readiness probes in the Pod specification.**

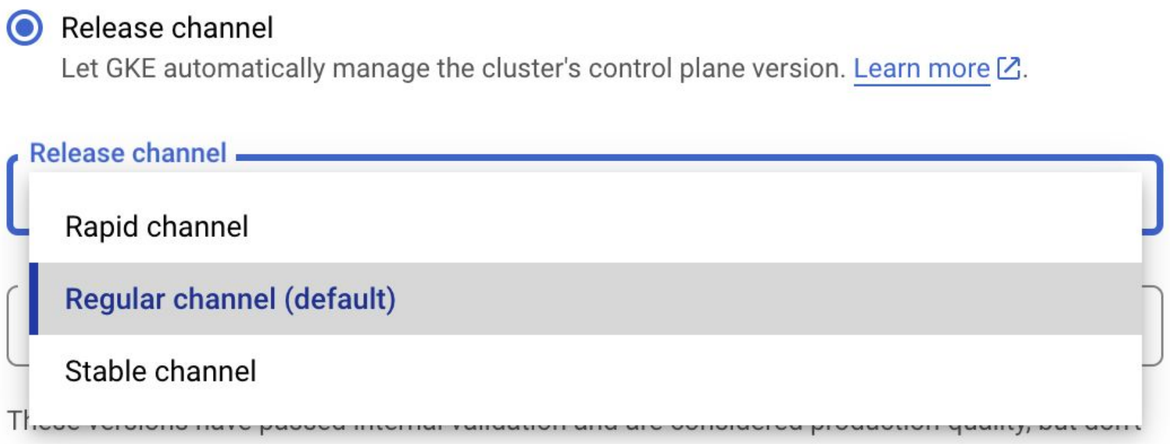
B. Configure health checks on the managed instance group.

C. Create a Scheduled Task to check whether the application is available.

D. Configure an uptime alert in Cloud Monitoring.

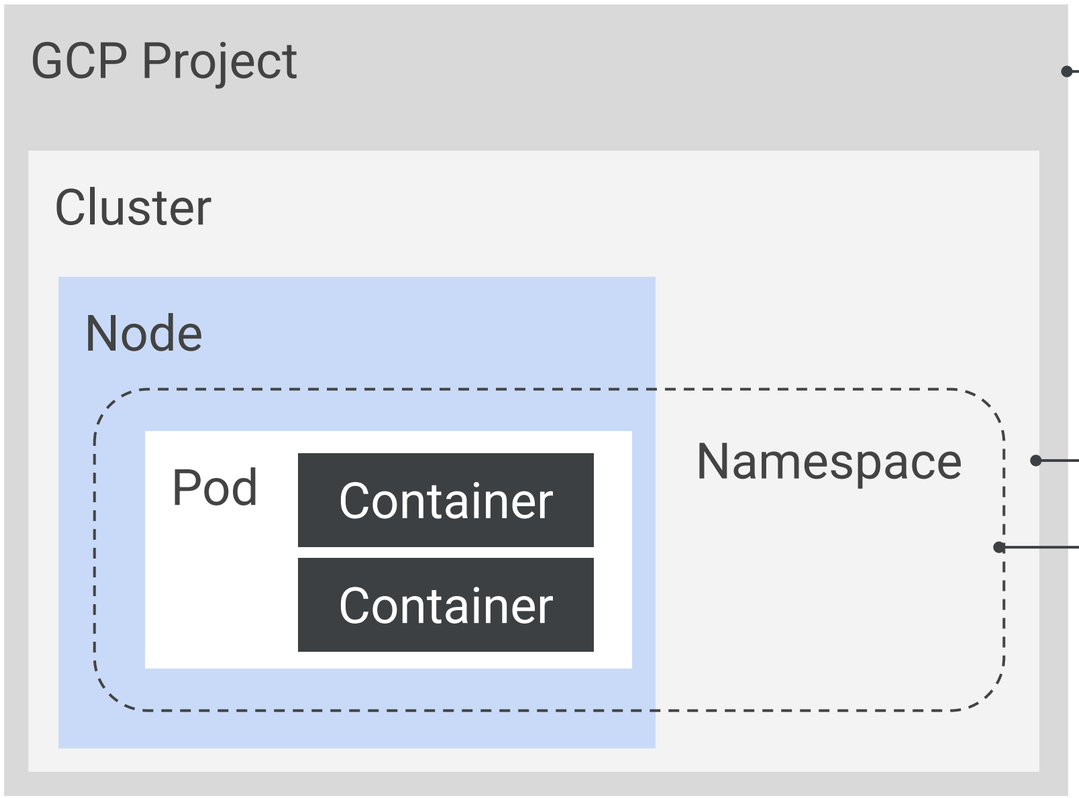
Best practices for GKE upgrades

1. **Setup multiple environments**: at a minimum pre-production and production clusters
2. **Enroll Clusters in Release Channels**: Stable or Regular release channels for production clust



1. **Create continuous upgrade strategy**: Receive updates about new GKE versions through cluster upgrade notifications through Pub/Sub
2. **Schedule maintenance windows and exclusions**: to increase upgrade predictability
3. **Set tolerance for disruption**: To ensure that pods sufficient number of replicas, use Pod Disruption Budget

GKE: Using IAM and RBAC



Use IAM at the project level
Set roles for

- Cluster Admin: manage clusters
- Container Developer: API access within clusters

Use RBAC at the cluster and namespace level
Set permissions on individual clusters and namespaces

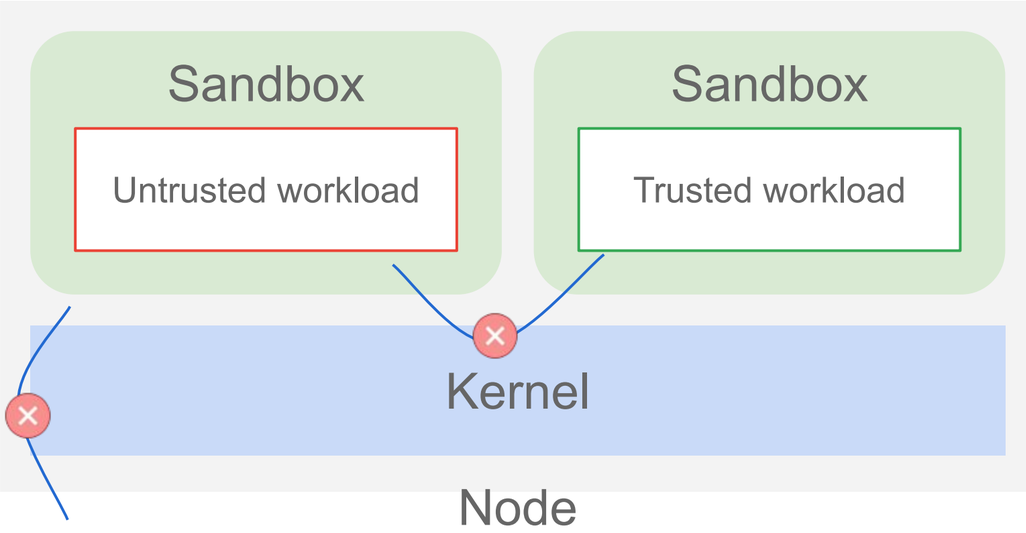
***Exam Tip**: IAM and Kubernetes RBAC work together to help manage access to your cluster. RBAC controls access on a cluster and namespace level, while IAM works on the project level*

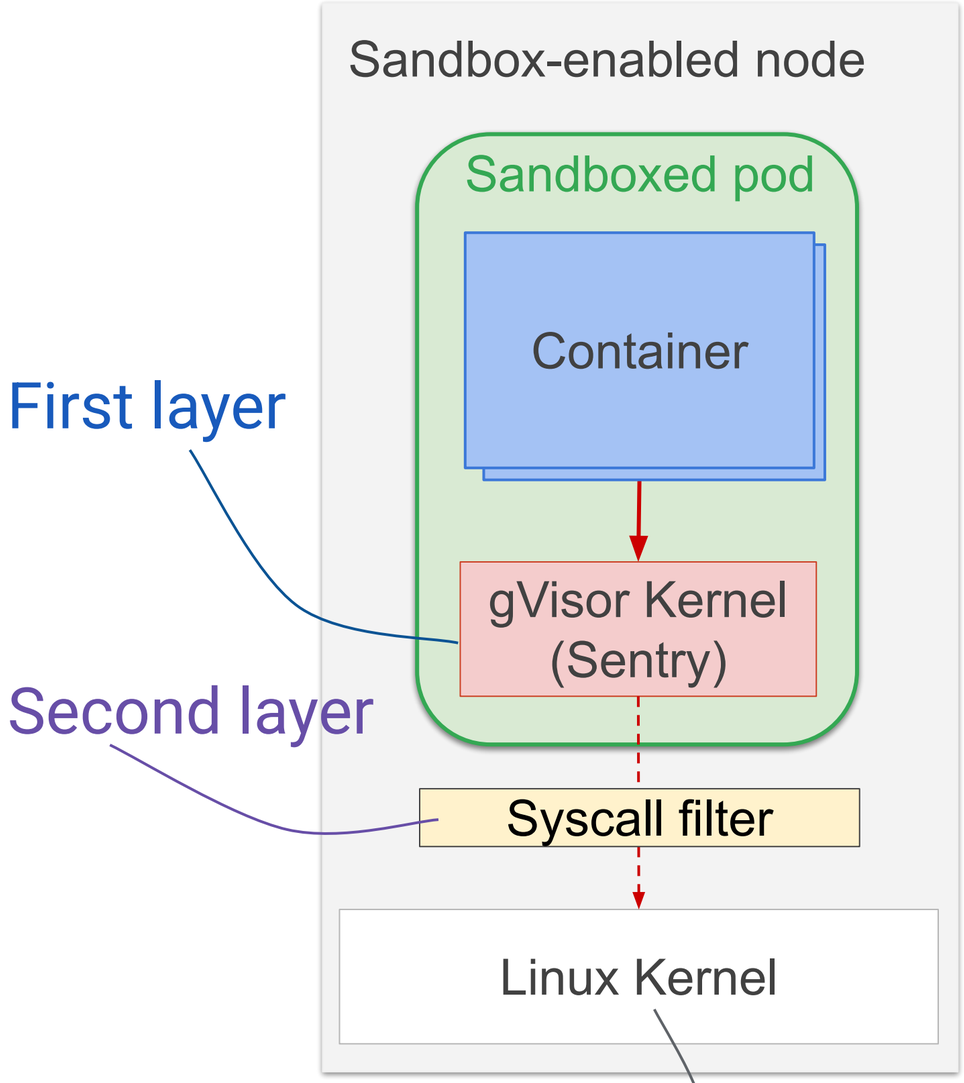
GKE Sandbox

Run **trusted and untrusted** workloads on the same node

Rather than achieving isolation via separate VMs, you can run workloads of different trust levels on the same node

Performance improvements from not having to allocate a new cluster to achieve isolation





***Exam Tip**: Commit 10 minutes to get an overview ofGKE Sandbox*

GKE networking: Subnet sizes

Subnet size for nodes	Maximum nodes	Maximum Pod IP addresses needed	Recommended Pod address range
/29	4	1,024	/21
/28	12	3,072	/20
/27	28	7,168	/19
/26	60	15,360	/18
/25	124	31,744	/17
/24	252	64,512	/16
/23	508	130,048	/15
/22	1,020	261,120	/14
/21	2,044	523,264	/13
/20	4,092	1,047,552	/12
/19	8,188	2,096,128	/11 (maximum Pod address range)

***Exam Tip**: make sure to watch this video to understand GKE networking well!*

GKE networking: Example

Subnet size for nodes	Maximum nodes	Maximum Pod IP addresses needed	Recommended Pod address range
/29	4	1,024	/21

$2^{(32-29)} = 8$ (4 of these are reserved for GCP)

110 pods running in each node -> **4 * 110 = 440**

Twice number of IPs per pod **440 * 2 = 880**

2^{10} number of IPs for pods

*Assuming the default maximum of 110 pods per node

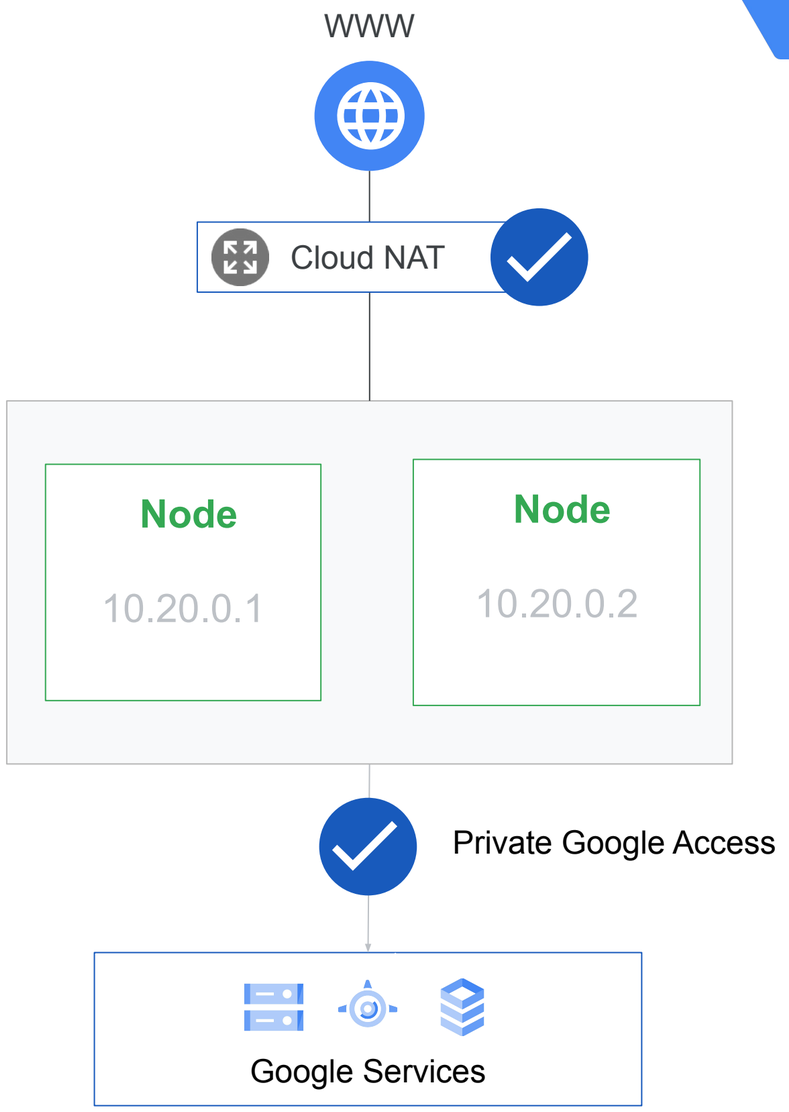
GKE best practices: Private Clusters

Private clusters isolate nodes from having inbound and outbound connectivity to the public internet

- Nodes have only private IP addresses
- Nodes use Private Google Access to communicate with Google APIs
- Nodes can use Cloud NAT to reach the internet
- Control Plane gets an additional private endpoint for the cluster nodes to talk to the control plane.

Exam Tip:

- *Private Clusters are definitely a best practice with GKE*
- *Having a Private Cluster does NOT mean you can't expose workloads via Services to the outside world!*



Diagnostic Question Discussion

Your team needs to create a Google Kubernetes Engine (GKE) cluster to host a newly built application that requires access to third-party services on the internet.
Your company does not allow any Compute Engine instance to have a public IP address on Google Cloud. You need to create a deployment strategy that adheres to these guidelines.

What should you do?

A. Configure the GKE cluster as a private cluster. Configure Private Google Access on the Virtual Private Cloud (VPC).

B. Configure the GKE cluster as a public cluster and then disable external IPs on each of the cluster nodes.

C. Configure the GKE cluster as a private cluster, and configure Cloud NAT Gateway for the cluster subnet.

D. Create a Compute Engine instance, and install a NAT Proxy on the instance. Configure all workloads on GKE to pass through this proxy to access third-party services on the Internet.

Diagnostic Question Discussion

Your team needs to create a Google Kubernetes Engine (GKE) cluster to host a newly built application that requires access to third-party services on the internet.
Your company does not allow any Compute Engine instance to have a public IP address on Google Cloud. You need to create a deployment strategy that adheres to these guidelines.

What should you do?

A. Configure the GKE cluster as a private cluster. Configure Private Google Access on the Virtual Private Cloud (VPC).

B. Configure the GKE cluster as a public cluster and then disable external IPs on each of the cluster nodes.

C. Configure the GKE cluster as a private cluster, and configure Cloud NAT Gateway for the cluster subnet.

D. Create a Compute Engine instance, and install a NAT Proxy on the instance. Configure all workloads on GKE to pass through this proxy to access third-party services on the Internet.

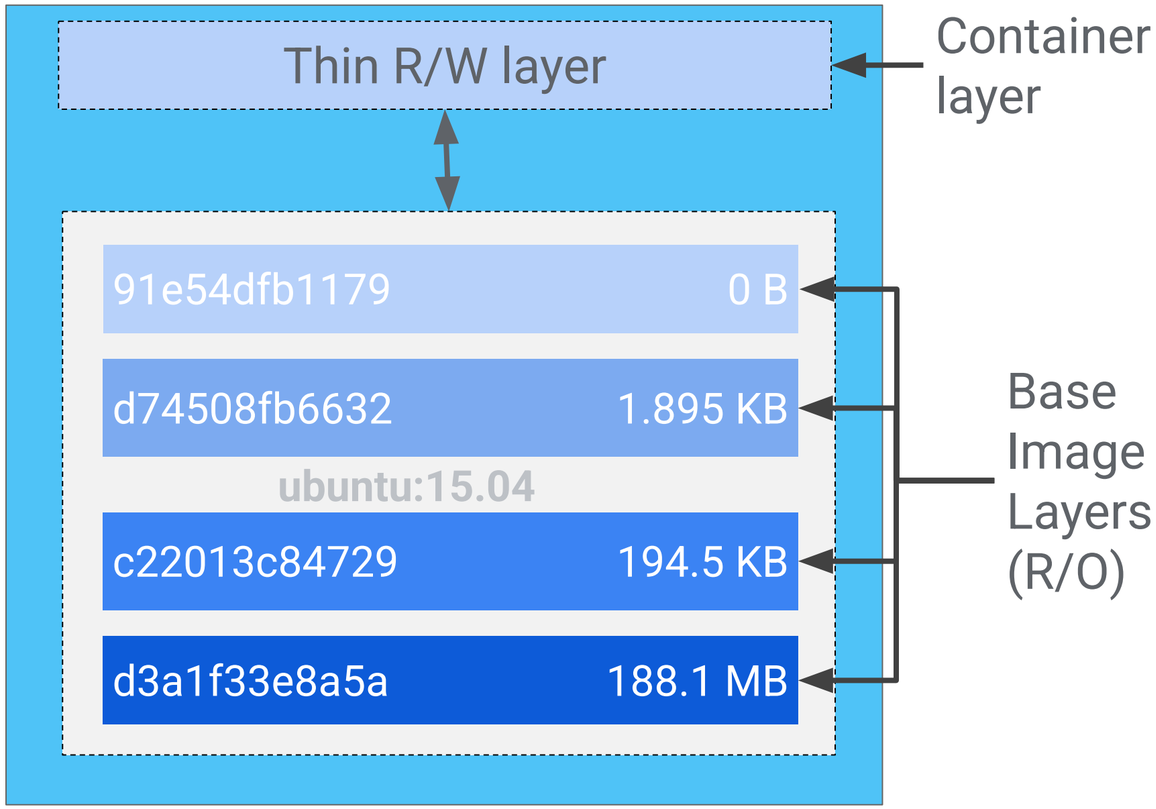
Container best practices: building images

Exam Tip: *Each instruction in a Dockerfile creates a layer.*

Dockerfile

```
FROM ubuntu:15.04
COPY . /app
RUN make /app
CMD python /app/app.py
```

Container



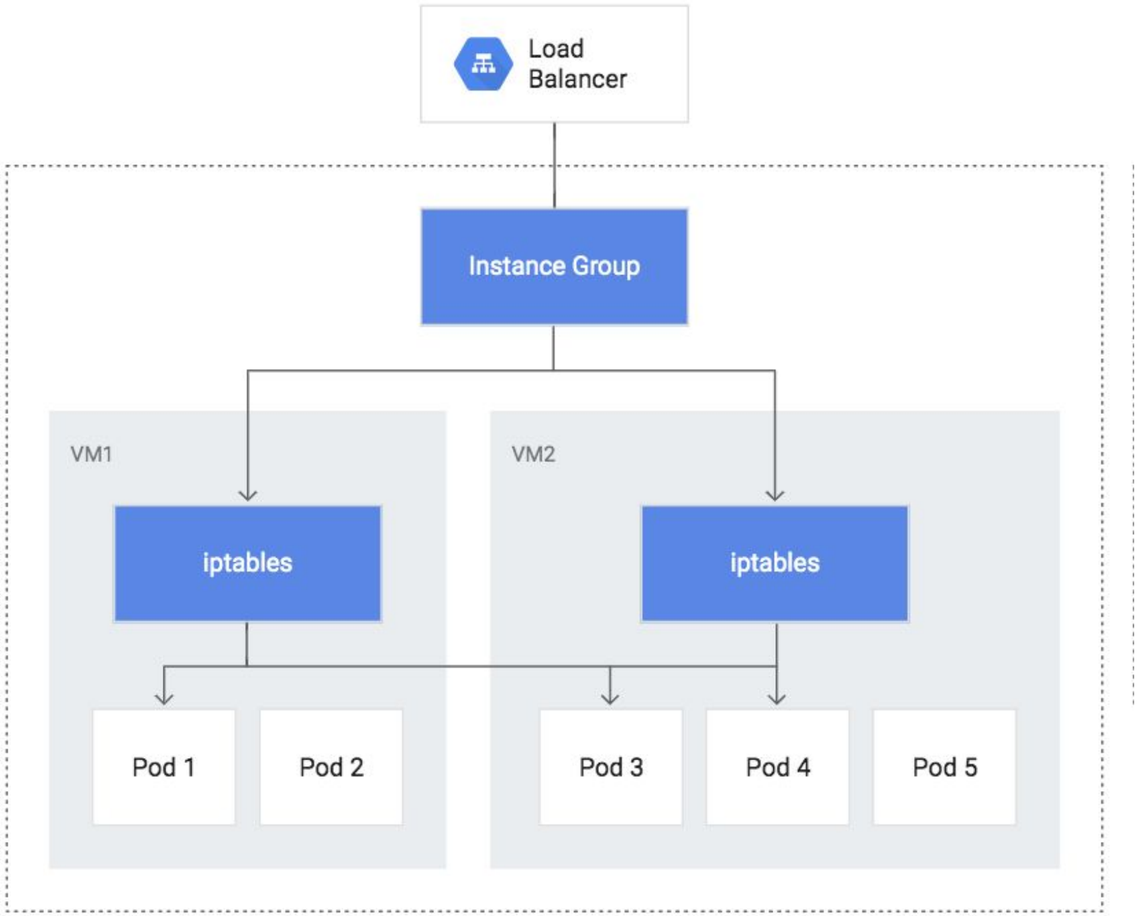
docker commands

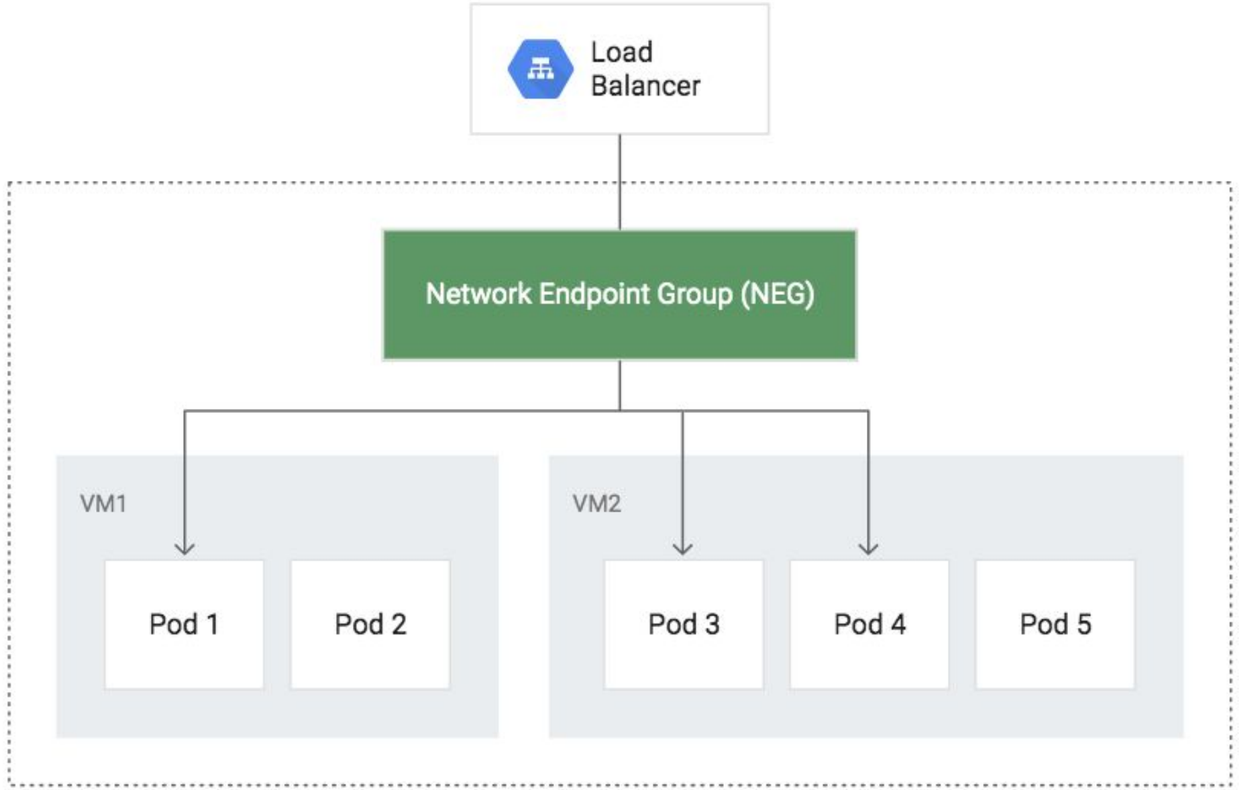
```
$> docker build -t py-web-server
.
$> docker run -d py-web-server
$> docker images
$> docker ps
$> docker logs <container id>
```

Exam Tips: *Here are best practices for building container images:*

- *Use the smallest base image possible (when new versions are rolled out, only smallest image layers are changed). Eg. use "alpine" image rather than "centos" or "ubuntu" if possible.*
- *Use multi-stage builds (app can be built in a first "build" container and the result can be used in another container)*
- *Try to create images with common layers (if a layer already exists on a cluster, it does not have to be downloaded)*

Ingress service: standard (non-NEG) vs NEG



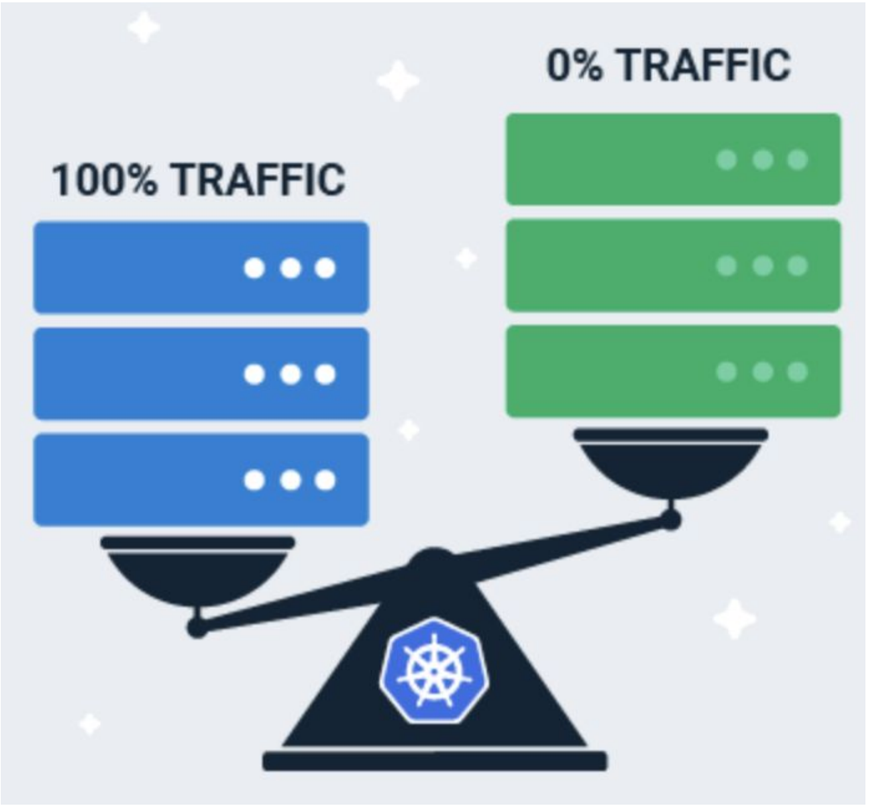


Exam Tip: *NEG is often preferred as a container-native load balancing type.*

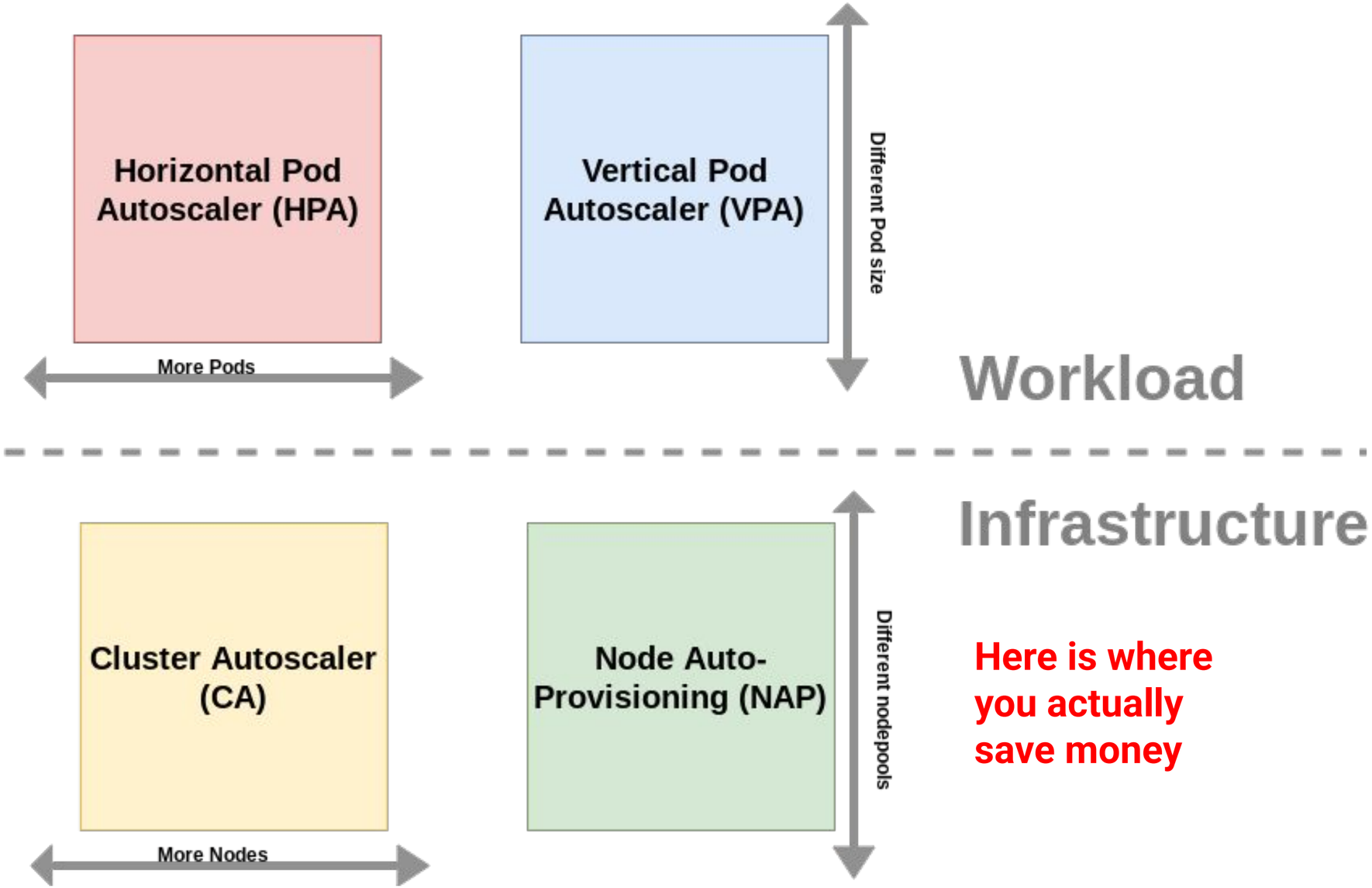
A/B testing, rolling updates, canary testing in GKE

Exam Tips:

- *You should know what deployment options GKE offers and how each of them works on a high level. Here is a great resource to understand those concepts.*
- *Differentiate between deployment strategies and testing strategies.*
- *Be able to choose the right strategy under different circumstances, eg. minimal downtime, rollback duration etc.*
- *Deploying new version is important… but being able to quickly and reliably roll back to previous version is even more important!*
- *To start a rolling update of a new app in GKE:*
  - *kubectl set image deployment/hello-app hello-app=REGION-docker.pkg.dev/${PROJECT_ID}/hello-repo/hello-app:v2*



GKE supports all 4 Kubernetes scalability dimensions



Diagnostic Question Discussion

Your company uses Google Kubernetes Engine (GKE) as a platform for all workloads. Your company has a single large GKE cluster that contains batch, stateful, and stateless workloads. The GKE cluster is configured with a single node pool with 200 nodes. Your company needs to reduce the cost of this cluster but does not want to compromise availability.

What should you do?

A. Create a second GKE cluster for the batch workloads only. Allocate the 200 original nodes across both clusters.

B. Configure CPU and memory limits on the namespaces in the cluster. Configure all Pods to have a CPU and memory limits.

C. Configure a HorizontalPodAutoscaler for all stateless workloads and for all compatible stateful workloads. Configure the cluster to use node auto scaling.

D. Change the node pool to use preemptible VMs.

Diagnostic Question Discussion

Your company uses Google Kubernetes Engine (GKE) as a platform for all workloads. Your company has a single large GKE cluster that contains batch, stateful, and stateless workloads. The GKE cluster is configured with a single node pool with 200 nodes. Your company needs to reduce the cost of this cluster but does not want to compromise availability.

What should you do?

A. Create a second GKE cluster for the batch workloads only. Allocate the 200 original nodes across both clusters.

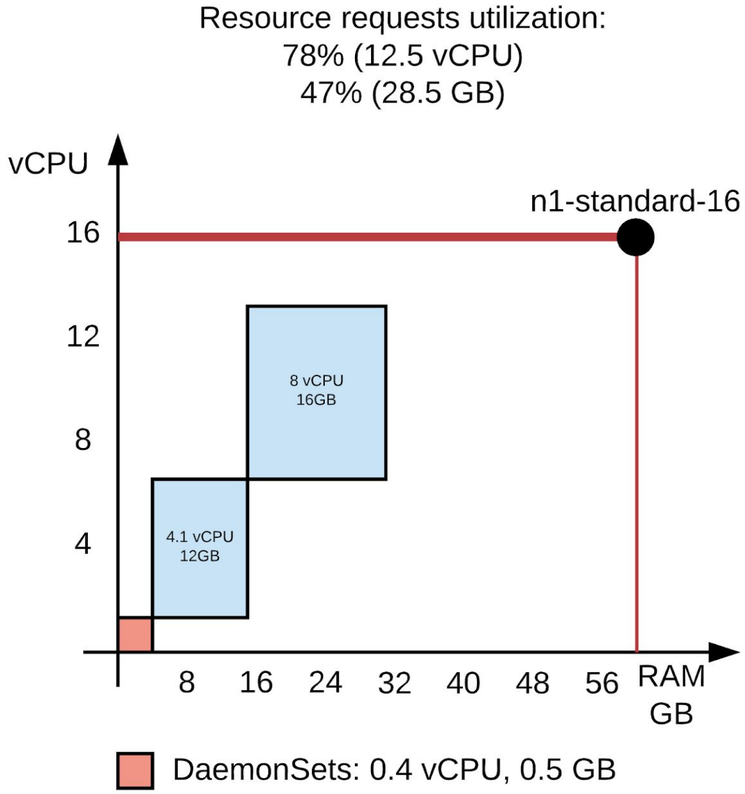
B. Configure CPU and memory limits on the namespaces in the cluster. Configure all Pods to have a CPU and memory limits.

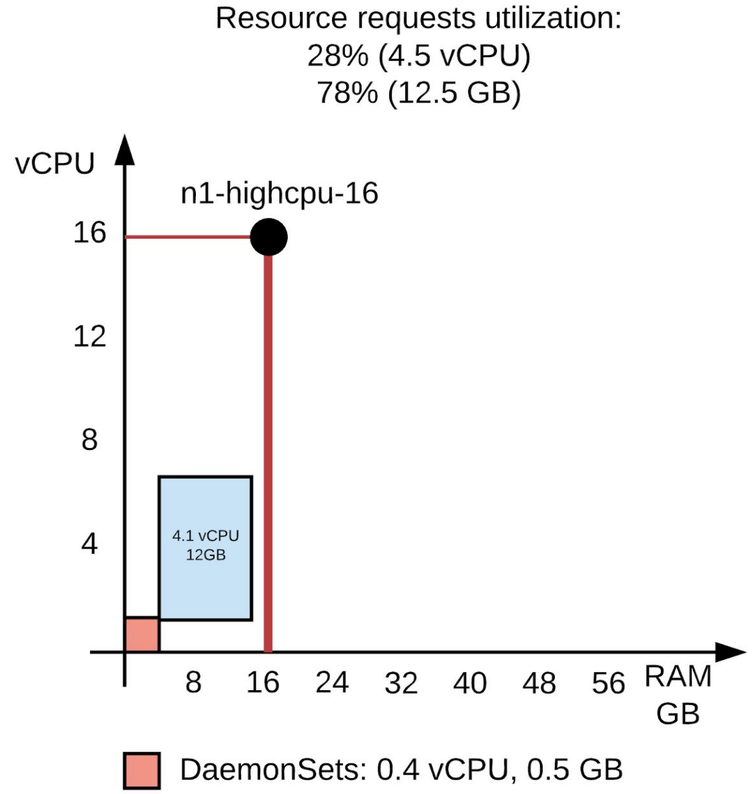
C. Configure a HorizontalPodAutoscaler for all stateless workloads and for all compatible stateful workloads. Configure the cluster to use node auto scaling.

D. Change the node pool to use preemptible VMs.

GKE: Binpacking

- Make sure your workload fit well inside the machine size
- You can create multiple node pools and use either nodeSelector or Node Affinity to select which node your pod must run.
- Another simpler option is to configure Node auto-provisioning





Diagnostic Question Discussion

You have deployed an application on Anthos clusters (formerly Anthos GKE). According to the SRE practices at your company, you need to be alerted if request latency is above a certain threshold for a specified amount of time.

What should you do?

A. Enable the Cloud Trace API on your project, and use Cloud Monitoring Alerts to send an alert based on the Cloud Trace metrics.

B. Install Anthos Service Mesh on your cluster. Use the Google Cloud Console to define a Service Level Objective (SLO), and create an alerting policy based on this SLO.

C. Use Cloud Profiler to follow up the request latency. Create a custom metric in Cloud Monitoring based on the results of Cloud Profiler, and create an Alerting policy in case this metric exceeds the threshold.

D. Configure Anthos Config Management on your cluster, and create a yaml file that defines the SLO and alerting policy you want to deploy in your cluster.

Diagnostic Question Discussion

You have deployed an application on Anthos clusters (formerly Anthos GKE). According to the SRE practices at your company, you need to be alerted if request latency is above a certain threshold for a specified amount of time.

What should you do?

A. Enable the Cloud Trace API on your project, and use Cloud Monitoring Alerts to send an alert based on the Cloud Trace metrics.

B. Install Anthos Service Mesh on your cluster. Use the Google Cloud Console to define a Service Level Objective (SLO), and create an alerting policy based on this SLO.

C. Use Cloud Profiler to follow up the request latency. Create a custom metric in Cloud Monitoring based on the results of Cloud Profiler, and create an Alerting policy in case this metric exceeds the threshold.

D. Configure Anthos Config Management on your cluster, and create a yaml file that defines the SLO and alerting policy you want to deploy in your cluster.

https://cloud.google.com/service-mesh/docs/observability/alert-policy-slo

GOOGLE Kubernetes Engine

#GCPSketchnote

@PVERGADIA THECLOUDGIRL.DEV

9.07.2020

What is GKE?

Your app crashed in production!

SYS ADMIN SAM

Let's fix it once and for all with CONTAINER!

But it worked on my machine?!

DEVELOPER ERIN

CONTAINER FEATURES

- Portable
- Shareable
- Versions
- Isolation
- Introspection
- Fast Deployments
- Immutable
- Reusable

CONTAINER	VIRTUAL MACHINE	BARE-METAL SERVER
Application Code	Application Code	Application Code
Dependencies	Dependencies	Dependencies
Operating System	Operating System	Operating System
Hardware	Hardware	Hardware

But I have got lots of containers, how can I orchestrate all of them?

SAM

I have heard Kubernetes could be the answer...

It is open source platform for managing containers.

ERIN

SAM

Kubernetes is not that easy from Installation to Provisioning to Upgrades to Scaling & SLAs

How does GKE work?

SAM

User

kubectl

http://gRPC/GCP UI

Google Kubernetes Engine Container Cluster in GCP

REGIONAL CONTROL PLANE
- API Server
- Resource Controllers
- Storage
- Scheduler
- Replica in Zone B
- Replica in Zone C

NODES
- User Pod: Containers
- User Pod: Containers
- User Pod: Containers

CONNECTED GCP SERVICES
- VPC Networking
- Persistent Disk
- Load Balancer
- Cloud Monitoring

Color Legend:
- Google Kubernetes Engine provisions, maintains & operates
- Google Kubernetes Engine provisions User optionally maintains & operates

MANAGED KUBERNETES

KUBERNETES → GKE

I bet GKE is the answer with advanced cluster management!

- Easy Cluster Creation
- Load Balancing
- Auto Scaling
- Auto Upgrades
- Auto Repair
- Logging Monitoring

How do I use GKE ?

Cloud Code → Code Repository → Cloud Build → Container Registry → GKE ← Kubectl

WORKLOAD		INFRASTRUCTURE	
Horizontal Pod Autoscaler (HPA)	Vertical Pod Autoscaler (VPA)	Cluster Autoscaler (CA)	Node Auto-Provisioning (NAP)
More Pods	Different Pod Sizes	More Nodes	Different Node Pools

How does it handle traffic growth?

How do I secure apps using GKE ?

SECURE APPS	TRUSTED NETWORKING	SOFTWARE SUPPLY CHAIN SECURITY
Data Encryption	Global VPC	Binary Authorization
Certified Images	Global Load Balancing	Vulnerability Scanning
Private Clusters	Cloud Armor	Managed Base Image
Idenity & Access - IAM - RBAC	Network Policy	

Cloud Run

Cloud Run - basics

- Enables stateless containers.
- Abstracts away infrastructure management.
- Automatically scales up and down.
- Open API and runtime environment.

Exam Tips:

- *“Stateless” is the key characteristic of Cloud Run.*
- *Cloud Run is MUCH newer than App Engine (2019 vs 2008) and uses Kubernetes (App Engine uses pre-K8s and pre-Docker containers). Otherwise, use-cases for App Engine and Cloud Run are similar.*

Containers in GCP = GKE or Cloud Run

OR

Exam Tip: *How to differentiate between GKE and Cloud Run?*

- *Cloud Run is fully serverless (GKE Standard was not… but Autopilot is…)*
- *Cloud Run are best when your biggest priority is time to market (fast development, deployment, scaling) and want to remove the ops and infra management from the process, or do not have a team to orchestrate and manage containers.*
- *98% of new Cloud Run users are able to code, build, and deploy an app within 5 minutes*

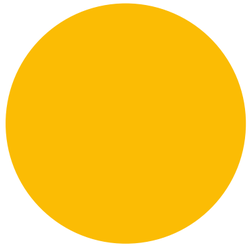
Cloud Run Functions

- Create single-purpose functions that respond to events without a server or runtime.
  - Event examples: New instance created, file added to Cloud Storage.
- Written in Javascript (Node.js), Python or Go; execute in managed Node.js environment on Google Cloud.

Exam Tip: *Cloud Run Functions can scale to 0 if not being actively used.*

Gen AI - Part 1

Evolution of AI Capabilities & Tools

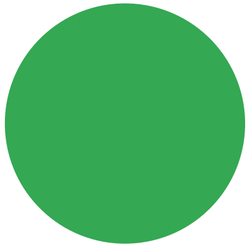


Predictive AI

Regression & Classification
Forecasting
Sentiment Analysis
Entity Extraction
Object Detection

Generative AI

Text, Image & Code Generation
Text & Code Rewriting & Formatting
Summarization
Extractive Q&A
Image & Video Descriptions

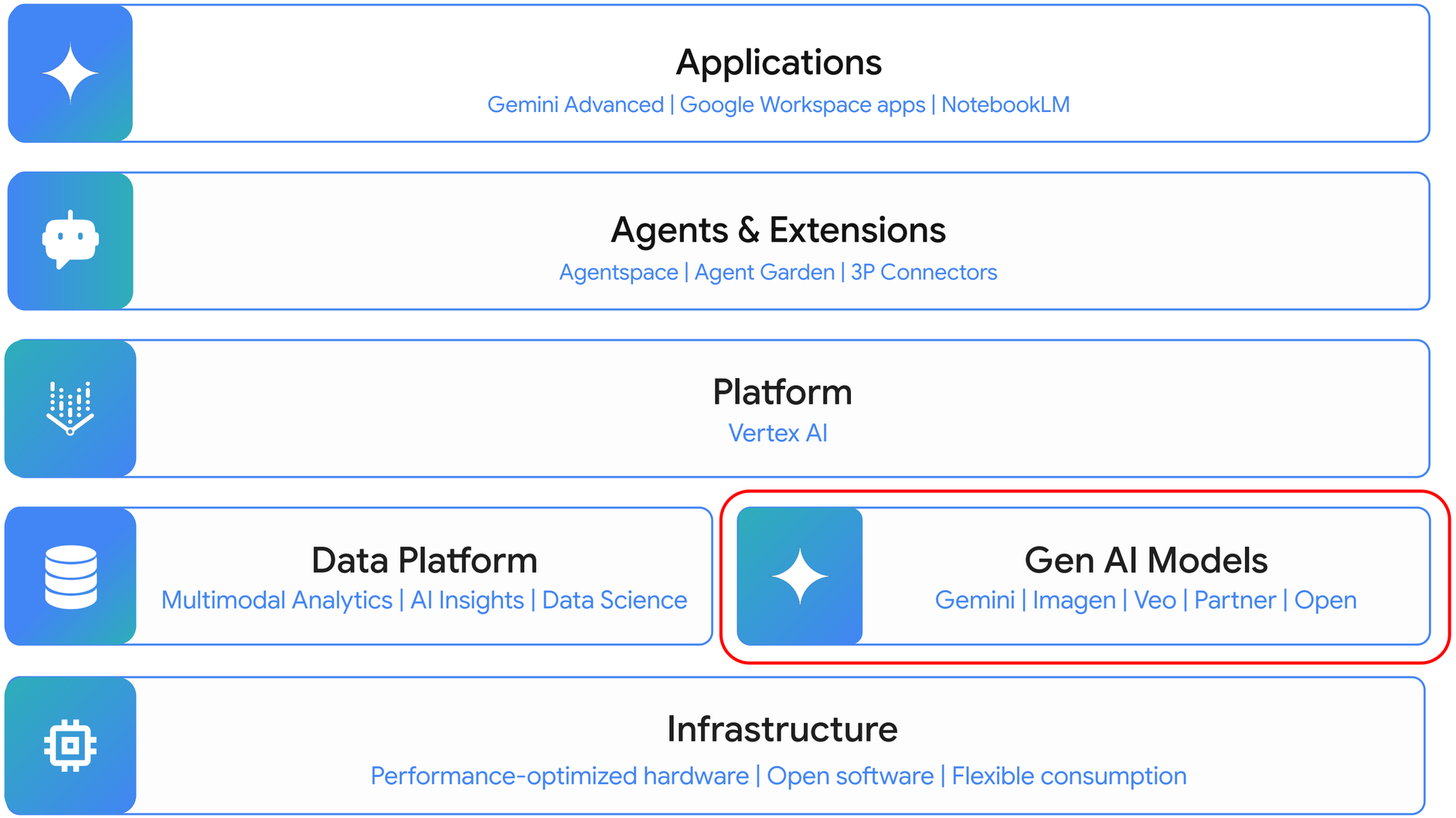


Multimodal Generative AI

Natural Image Understanding
Video Question Answering
Automatic Speech Recognition & Translation
Spatial Reasoning and Logic
Mathematical Reasoning in Visual Contexts

Train | Serve | MLOps | Prompt | Tune | RAG

Google's Unified AI technology stack



Google Gen AI Models

Across a variety of model sizes to address use cases

Gemini (Pro / Flash / Flash-Lite)

Sophisticated multimodal reasoning across multiple tasks and domains

Gemma family

Lightweight open models

Imagen for Text to Image

Create and edit images from simple prompts

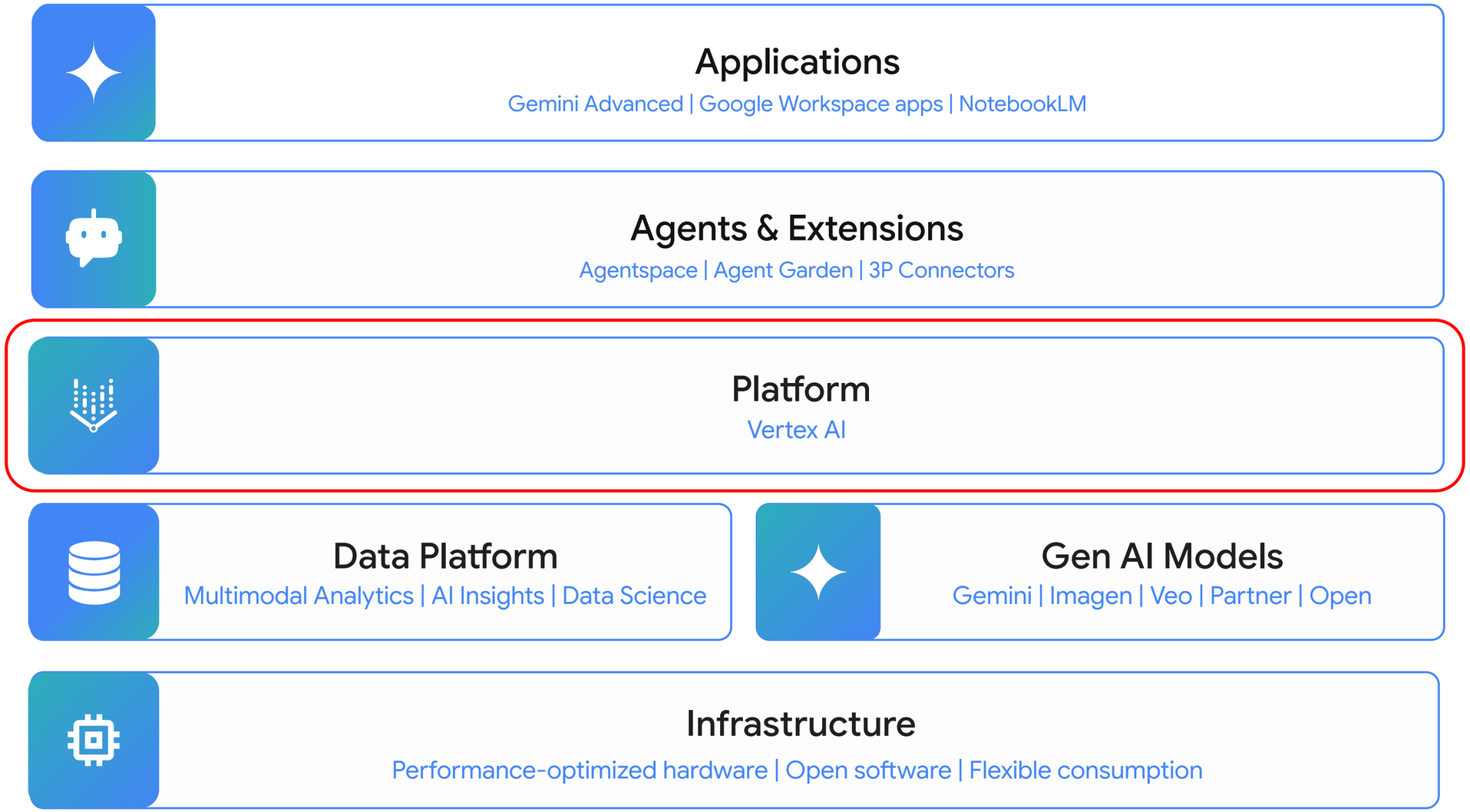
Embedding Models for Text and Image

Extract semantic information from unstructured data

Veo models

Text and images to generate novel videos

Google's Unified AI technology stack



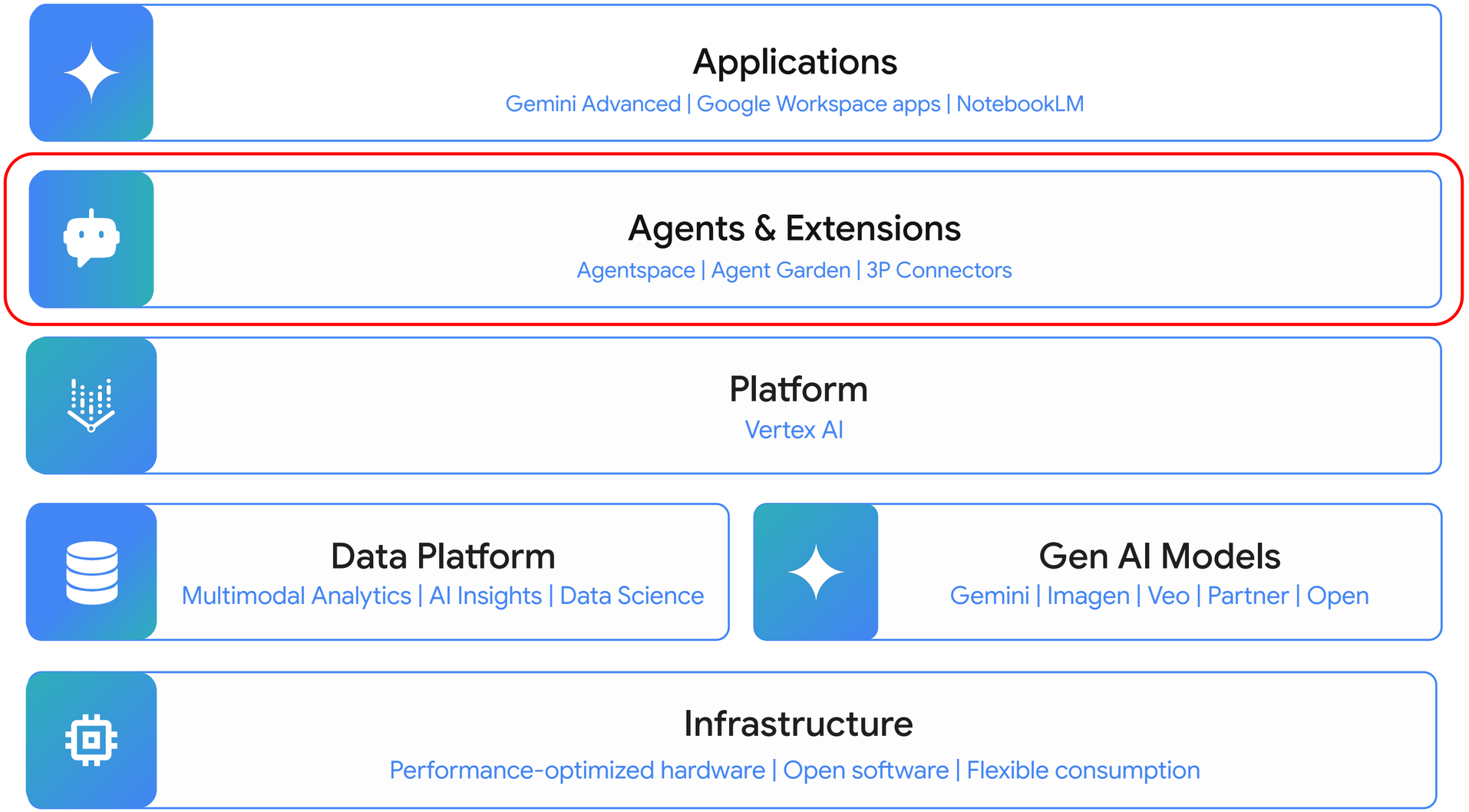
What is Vertex AI?

Vertex AI

cloud.google.com/vertex-ai

- Managed, End-to-End **AI & ML Platform on Google Cloud**
- **Model Garden** - Generative & Predictive AI Models from Google, Partners and Open Source
- **Vertex AI Studio** - Experiment with Models
- **Custom Models** - Training/Prediction Pipelines
- **Vector Search** for Embeddings
- **Colab Enterprise** for Jupyter Notebooks
- **Enterprise-Grade** Security/Reliability

Google's Unified AI technology stack



Observe, Act, Achieve

An AI Agent is an **application** that tries to achieve a **goal** by **observing** the world and **acting** upon it using the **tools** it has at its disposal.

AI agents: The next frontier of software

Autonomous action

Agents can perform complex **tasks** and workflows with minimal human intervention.

Reasoning and planning

Agents leverage advanced AI models to make informed decisions and **adapt** to changing environments.

Continuous learning

Agents can **learn** from experience and improve their performance over time.

Multi-agent collaboration

Agents can work **together** to achieve shared goals, unlocking new levels of complexity and efficiency.

Chatbots → **Indispensable tools for getting work done**

Google Cloud agent strategy

Build your own agents

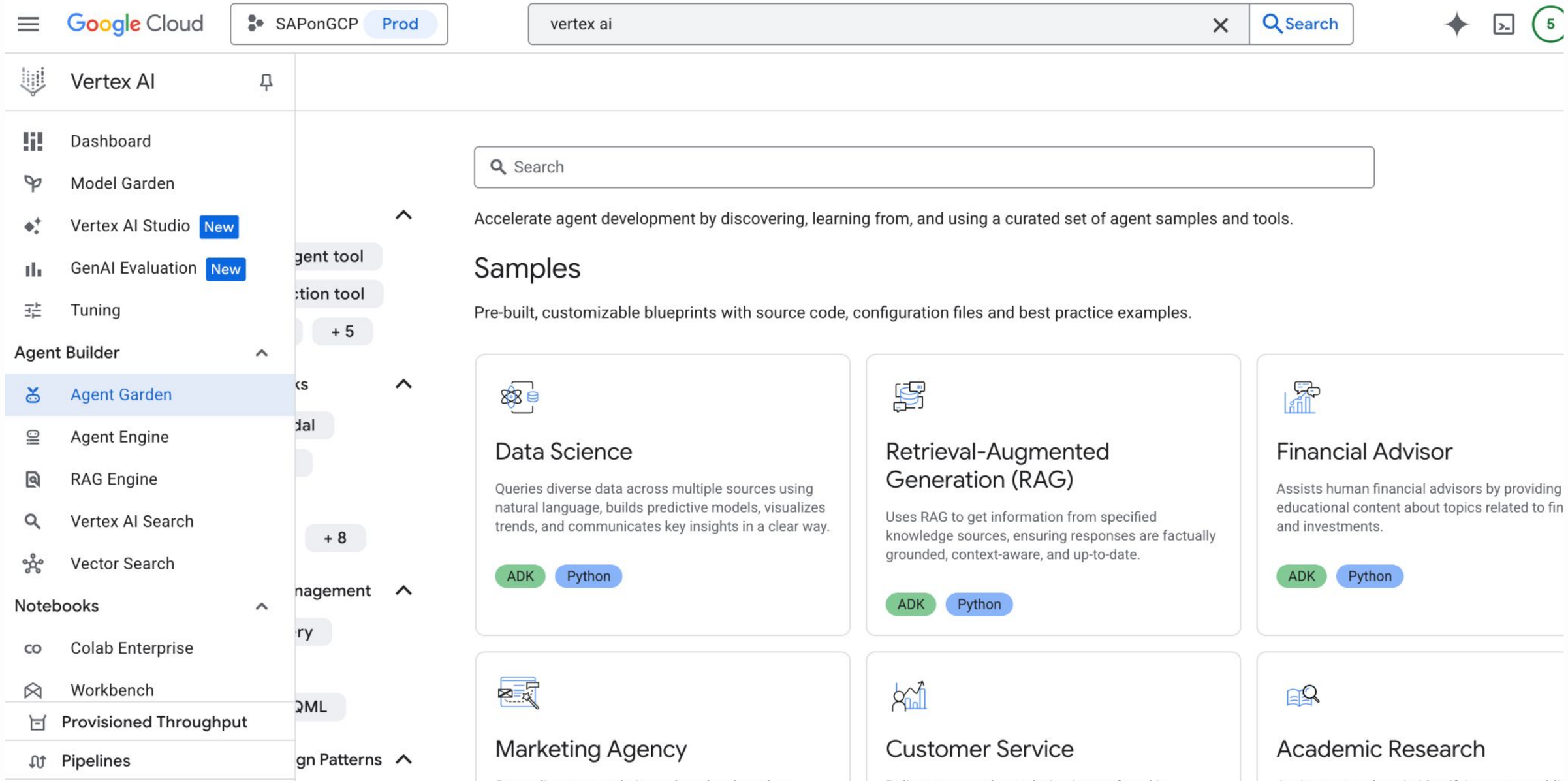
Use Google Cloud agents

Bring in partner agents

Enable interoperability with Model Context Protocol + Agent2Agent Protocol

Gemini Enterprise

Use Google Cloud's ready-packaged agents



I want to build an Agent myself

Which Google Cloud option should I use?

Less effort — More effort

Simpler Agents — More flexibility & control

Gemini Enterprise
(Employee Productivity)

No code UI to build Agents focused on internal employee productivity use cases

Conversational Agents
(Customer Engagement)

No code UI to build Agents focused on customer experience use cases

ADK + Vertex AI Agent Engine

Use **Agent Development Kit** to create your agents and deploy on **Vertex AI Agent Engine**

OSS or from scratch + Vertex AI Agent Engine

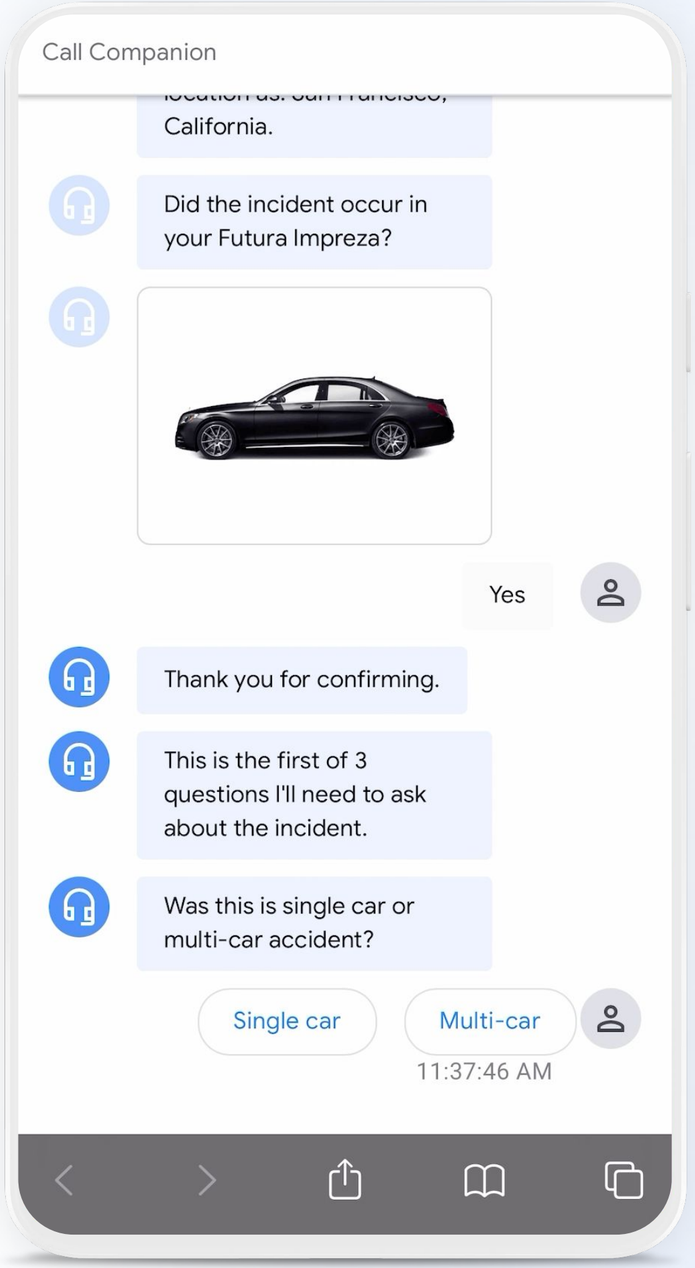
Use any agent framework like LangGraph, AG2, CrewAI, etc, and easily deploy, autoscale, manage, and add session, memory, and more services with **Vertex AI Agent Engine**

Customer Engagement Suite with Google AI

Conversational Agents

Instantly resolve inquiries with AI-powered Conversational Agent

- Proactive, personalized 24/7 self-service.
- Deploy complex AI agents in clicks with a no-code console.
- Rich, multimodal interactions with voice, text, and images

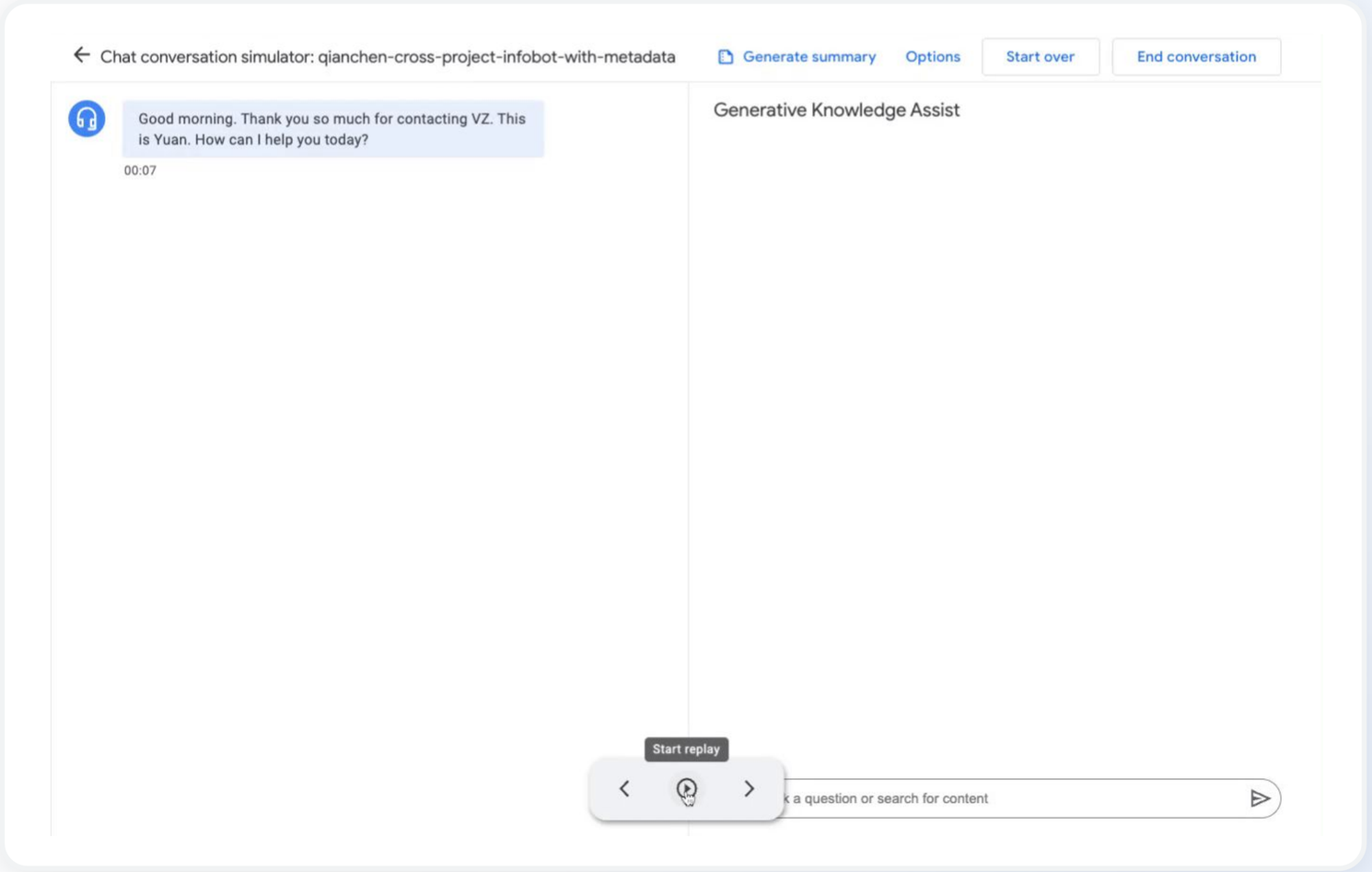


Google Cloud

Agent Assist

Supercharge Customer Support Representatives with AI assistance

- Increase agent productivity and customer satisfaction.
- Guide agents through complex issues with real-time coaching.
- Automate call summaries to reduce after-call work.



Google Cloud

Conversational Insights and Quality AI

Turn customer conversations into business-wide intelligence

- Automatically review every conversation for quality.
- Identify top customer issues and agent coaching needs.
- Use customer feedback to improve products and services.

Customer Engagement Suite

The end-to-end platform for transforming your customer experiences

Conversational Agents

Intelligent, instant 24/7 self-service

Agent Assist

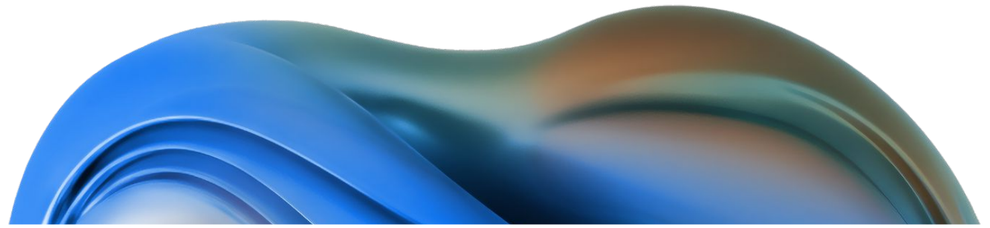
AI-powered coaching and next-best action guidance

Conversational Insights and Quality AI

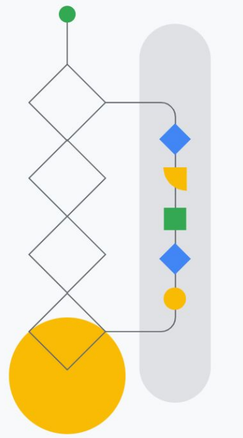
Gain insights to improve products and services

Contact Center as a Service

Secure, scalable, and omnichannel foundation



The Generative AI Landscape:
Workflow Agents



Workflow agents are designed to streamline your work and make sure things get done efficiently and correctly by automating tasks or going through complex processes.

- **You provide input:** You define a task or trigger a process like submitting a form, uploading a file, initiating a scheduled event, or even ordering a product online.
- **The agent understands:** The agent is the software that automates those steps. It interprets the task's requirements and defines the series of steps needed to complete the task.
- **The agent calls a tool:** Based on the workflow's definition, the agent executes a series of actions. This could involve data transformation, file transfer, sending notifications, integrating with external systems, or initiating other automated processes using APIs.
- **The agent generates a result/output:** It compiles the outcome of the executed actions, which might be a report, a data file, a confirmation message, or an updated status within a system.
- **The agent delivers the result/output:** The agent delivers the output to the designated recipient(s) or system(s), such as via email, a dashboard, a database update, or a file storage location.

- **Ecommerce order fulfillment:** An agent automatically processes orders, updates inventory, sends shipping notifications, and handles returns.
- **Customer onboarding:** An agent guides new customers through account setup, provides tutorials, and answers frequently asked questions.
- **Automated research:** An agent can conduct in-depth research on a given topic by autonomously browsing the web, summarizing relevant content, and generating comprehensive reports. (Try this out with Gemini Deep Research.)
- **Security Log Parsing:** An agent that inspects incoming security logs for abnormalities and can flag them, open a ticket, begin triage, and assign to humans for review when necessary.

Altostrat Media

case study

Proposed Technical Solution

- Altostrat currently utilizes GKE for content management and delivery, and Cloud Run for serverless tasks like video transcoding and metadata extraction
    - Implement Istio / Anthos / ASM / Cloud Service Mesh / Fleets to provide a centralized management platform for both Google Cloud and the on-premises legacy systems
    - See more about container orchestration
    - Continue using Cloud Run functions for event-driven tasks that require serverless execution, such as video transcoding and personalized content recommendations, which scales with minimal latency
- Use Cloud Storage Lifecycle Management to optimize costs for growing media library (documents, audio, video).
    - Use Storage Transfer Service to migrate large-scale on-premises archival data (over 1 TB) to Cloud Storage
- Real-time Processing and Data Flow:
    - Integrate Pub/Sub with Dataflow to create streaming pipelines for real-time parallel data processing, preparing data for analysis in BigQuery
- Continue leveraging BigQuery in combination with BI tools (like Looker or Tableau) for interactive data exploration and decision-making on content strategy
- AI-related:
    - **Content Enrichment and Metadata Extraction:** Use pre-trained AI services such as Video Intelligence API and Natural Language API to automatically extract rich metadata from media assets, enabling content discovery, dynamic pricing, and targeted marketing
    - **Harmful Content Detection:** use features such as Model Armor, use Vertex AI content filters and / or develop custom AI-powered detection.
    - **Generative AI for User Experience and Content Virality:** Build AI chatbots leveraging LLMs and Conversational AI (e.g., Dialogflow or specific Vertex AI features). Also, implement Generative AI to automatically generate concise summaries of diverse media content
    - **Model Management and Auditing:** Utilize Vertex AI functionalities such as: Model evaluation, Explainable AI, Vertex AI Workbench or Colab Notebooks, Model Monitoring, Vertex AI Model Registry etc.
- Plus others: Cloud Build for CI/CD modernization, Hybrid Connectivity options, use Google Cloud Observability platform etc

[Altostrat Media case study] Diagnostic Question #1

Altostrat Media stores a vast and growing library of video content in Cloud Storage. The majority of their archived documentaries (which comprise 60% of their total volume) are accessed less than four times per year, and many are retained for long-term regulatory purposes. Altostrat needs to immediately optimize cloud storage costs for these growing media volumes while ensuring long-term retention requirements are met and maintaining high availability for serving the content globally

Which storage solution best balances cost optimization, global availability, and long-term regulatory compliance?

A. Use a Multi-Regional Cloud Storage bucket and apply an Object Lifecycle Management policy to transition objects accessed less than once a year to Archive Storage.

B. Use a Regional Coldline Storage bucket, and rely on Object Versioning for long-term retention assurance.

C. Store all content in a Multi-Regional Standard Storage bucket to ensure high availability, and manually move documentary assets to Archive Storage only after 12 months using a scheduled data job.

D. Use a Dual-Region Nearline Storage bucket, ensuring objects are moved to Coldline Storage after 90 days of inactivity using Object Lifecycle Management

[Altostrat Media case study] Diagnostic Question #1

Altostrat Media stores a vast and growing library of video content in Cloud Storage. The majority of their archived documentaries (which comprise 60% of their total volume) are accessed less than four times per year, and many are retained for long-term regulatory purposes. Altostrat needs to immediately optimize cloud storage costs for these growing media volumes while ensuring long-term retention requirements are met and maintaining high availability for serving the content globally

Which storage solution best balances cost optimization, global availability, and long-term regulatory compliance?

A. Use a Multi-Regional Cloud Storage bucket and apply an Object Lifecycle Management policy to transition objects accessed less than once a year to Archive Storage.

B. Use a Regional Coldline Storage bucket, and rely on Object Versioning for long-term retention assurance.

C. Store all content in a Multi-Regional Standard Storage bucket to ensure high availability, and manually move documentary assets to Archive Storage only after 12 months using a scheduled data job.

D. Use a Dual-Region Nearline Storage bucket, ensuring objects are moved to Coldline Storage after 90 days of inactivity using Object Lifecycle Management

[Altostrat Media case study] Diagnostic Question #2

Altostrat Media aimed for enhanced reach and personalization. From an architectural perspective, how would you design a solution on GCP to dynamically segment audiences and deliver personalized ad content at scale, considering both batch and real-time data processing needs?

A. Utilize Dataflow for ETL from disparate sources into Cloud SQL, then integrate with Marketing Platform.

B. Ingest real-time events via Pub/Sub to Dataflow for stream processing and feature engineering, storing results in BigQuery for ML model training with Vertex AI, and serving predictions via custom APIs on Cloud Run.

C. Store all data in Cloud Storage buckets, use Dataproc for occasional batch processing, and manually update ad campaigns.

D. Implement App Engine to host custom audience segmentation logic and connect directly to ad networks.

[Altostrat Media case study] Diagnostic Question #2

Altostrat Media aimed for enhanced reach and personalization. From an architectural perspective, how would you design a solution on GCP to dynamically segment audiences and deliver personalized ad content at scale, considering both batch and real-time data processing needs?

A. Utilize Dataflow for ETL from disparate sources into Cloud SQL, then integrate with Marketing Platform.

B. Ingest real-time events via Pub/Sub to Dataflow for stream processing and feature engineering, storing results in BigQuery for ML model training with Vertex AI, and serving predictions via custom APIs on Cloud Run.

C. Store all data in Cloud Storage buckets, use Dataproc for occasional batch processing, and manually update ad campaigns.

D. Implement App Engine to host custom audience segmentation logic and connect directly to ad networks.

[Altostrat Media case study] Diagnostic Question #3

A key business requirement is to enable natural language interaction with the platform and provide 24/7 personalized user support via advanced chatbots. This chatbot needs to utilize Natural Language Understanding (NLU) to answer complex queries about content and suggest personalized recommendations.

Which Google Cloud service combination should be used to build and deploy this critical user engagement component?

A. Use a custom Python script deployed on a managed instance group (MIG) for NLU, and integrate it with BigQuery ML for recommendations.

B. Use a monolithic Node.js application deployed on GKE to manage all user interactions and recommendation logic

C. Utilize the Natural Language API directly within Cloud Functions to analyze user text input, avoiding any need for stateful conversational platforms

D. Leverage Conversational AI (such as Dialogflow or Vertex AI's NLU capabilities) to handle user interaction and intent recognition

[Altostrat Media case study] Diagnostic Question #3

A key business requirement is to enable natural language interaction with the platform and provide 24/7 personalized user support via advanced chatbots. This chatbot needs to utilize Natural Language Understanding (NLU) to answer complex queries about content and suggest personalized recommendations.

Which Google Cloud service combination should be used to build and deploy this critical user engagement component?

A. Use a custom Python script deployed on a managed instance group (MIG) for NLU, and integrate it with BigQuery ML for recommendations.

B. Use a monolithic Node.js application deployed on GKE to manage all user interactions and recommendation logic

C. Utilize the Natural Language API directly within Cloud Functions to analyze user text input, avoiding any need for stateful conversational platforms

D. Leverage Conversational AI (such as Dialogflow or Vertex AI's NLU capabilities) to handle user interaction and intent recognition

[Altostrat Media case study] Diagnostic Question #4

Altostrat currently relies on Cloud Monitoring and Prometheus, but alerts for critical system issues are primarily delivered via email. To accelerate and enhance the reliability of operational workflows, the architecture team needs to implement a low-latency, highly reliable alerting mechanism for major issues (e.g., GKE cluster failure or high-priority transcoding errors)

What action should the architect take to improve the immediacy and reliability of critical alerts?

A. Set up a dedicated Compute Engine instance to continuously parse incoming email alerts and manually execute runbooks.

B. Increase the retention period of Prometheus metrics to allow for better post-mortem analysis

C. Configure Cloud Monitoring to integrate with an external service like PagerDuty or Slack via Notification Channels

D. Use Cloud Deployment Manager to standardize the deployment of existing Cloud Monitoring dashboards

[Altostrat Media case study] Diagnostic Question #4

Altostrat currently relies on Cloud Monitoring and Prometheus, but alerts for critical system issues are primarily delivered via email. To accelerate and enhance the reliability of operational workflows, the architecture team needs to implement a low-latency, highly reliable alerting mechanism for major issues (e.g., GKE cluster failure or high-priority transcoding errors)

What action should the architect take to improve the immediacy and reliability of critical alerts?

A. Set up a dedicated Compute Engine instance to continuously parse incoming email alerts and manually execute runbooks.

B. Increase the retention period of Prometheus metrics to allow for better post-mortem analysis

C. Configure Cloud Monitoring to integrate with an external service like PagerDuty or Slack via Notification Channels

D. Use Cloud Deployment Manager to standardize the deployment of existing Cloud Monitoring dashboards

[Altostrat Media case study] Diagnostic Question #5

Altostrat deploys new applications with stateful information (like user management data) on GKE and uses Cloud SQL for the managed relational database backend. To ensure low latency and improved network security, the Cloud SQL instance is configured with a Private IP.

Which is the most secure and recommended way to ensure the GKE pods can connect reliably to the Cloud SQL instance?

A. Enable a Public IP address on the Cloud SQL instance and use SSL certificates for connection

B. Set up a dedicated VM as a jump host within the VPC network to proxy all database connections

C. Use the Cloud SQL Auth Proxy running as a sidecar container within the GKE pods to manage secure, IAM-based connectivity over the Private Service Access network

D. Configure the Cloud SQL instance with an Authorized Network (CIDR block) that encompasses the entire GKE node IP range

[Altostrat Media case study] Diagnostic Question #5

Altostrat deploys new applications with stateful information (like user management data) on GKE and uses Cloud SQL for the managed relational database backend. To ensure low latency and improved network security, the Cloud SQL instance is configured with a Private IP.

Which is the most secure and recommended way to ensure the GKE pods can connect reliably to the Cloud SQL instance?

A. Enable a Public IP address on the Cloud SQL instance and use SSL certificates for connection

B. Set up a dedicated VM as a jump host within the VPC network to proxy all database connections

C. Use the Cloud SQL Auth Proxy running as a sidecar container within the GKE pods to manage secure, IAM-based connectivity over the Private Service Access network

D. Configure the Cloud SQL instance with an Authorized Network (CIDR block) that encompasses the entire GKE node IP range

Optional materials 1

[VIDEOS]

- How is data encrypted? How does encryption work at Google's data centers?
- Data Encryption and KMS: Data Encryption and Managed Encryption Keys
- What is Kubernetes?
- Cloud Run intro: Say hello to serverless containers with Cloud Run
- VERY nice Cloud Run deep-dive session: How to run your container without servers
- Examples of Cloud Run usage: Can Cloud Run handle these 9 workloads?
- Where should I run my code?:
  a. Shorter version: Choosing the right compute option in GCP: a decision tree
  b. Longer version (HIGHLY recommended!): Where should I run my stuff? Choosing compute options

Make sure to…

Enjoy the journey as much as the destination!