O'REILLY®

Google Cloud Professional Cloud Architect Crash Course

Day 1: Course Schedule

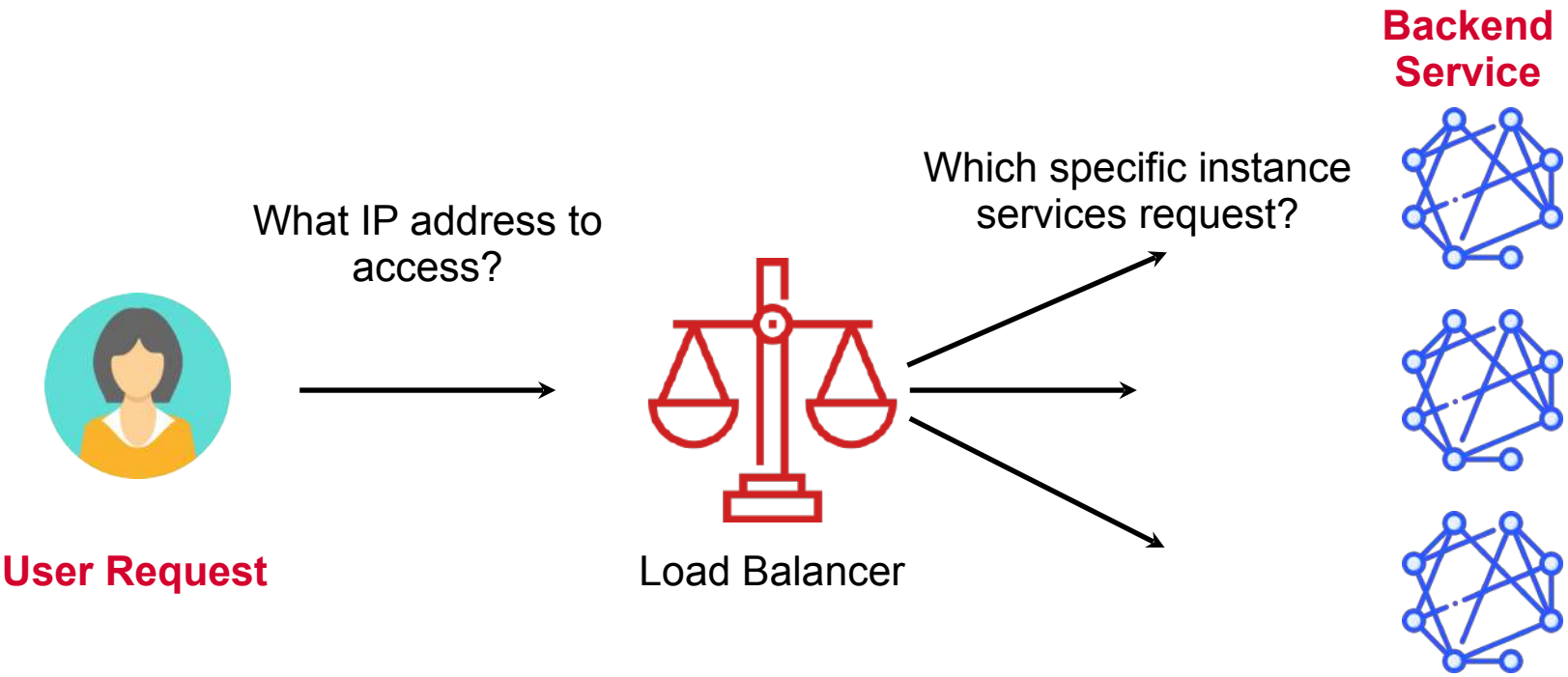
- Big Picture: Terms and concepts
- Resource Hierarchy on Google Cloud
- Infrastructure-as-a-Service
  - Google Compute Engine
  - Managed Instance Groups
- Serverless Compute
  - Cloud Run
  - Cloud Run Functions
  - App Engine
- Google Kubernetes Engine (GKE)
- Load Balancing

Day 2: Course Schedule

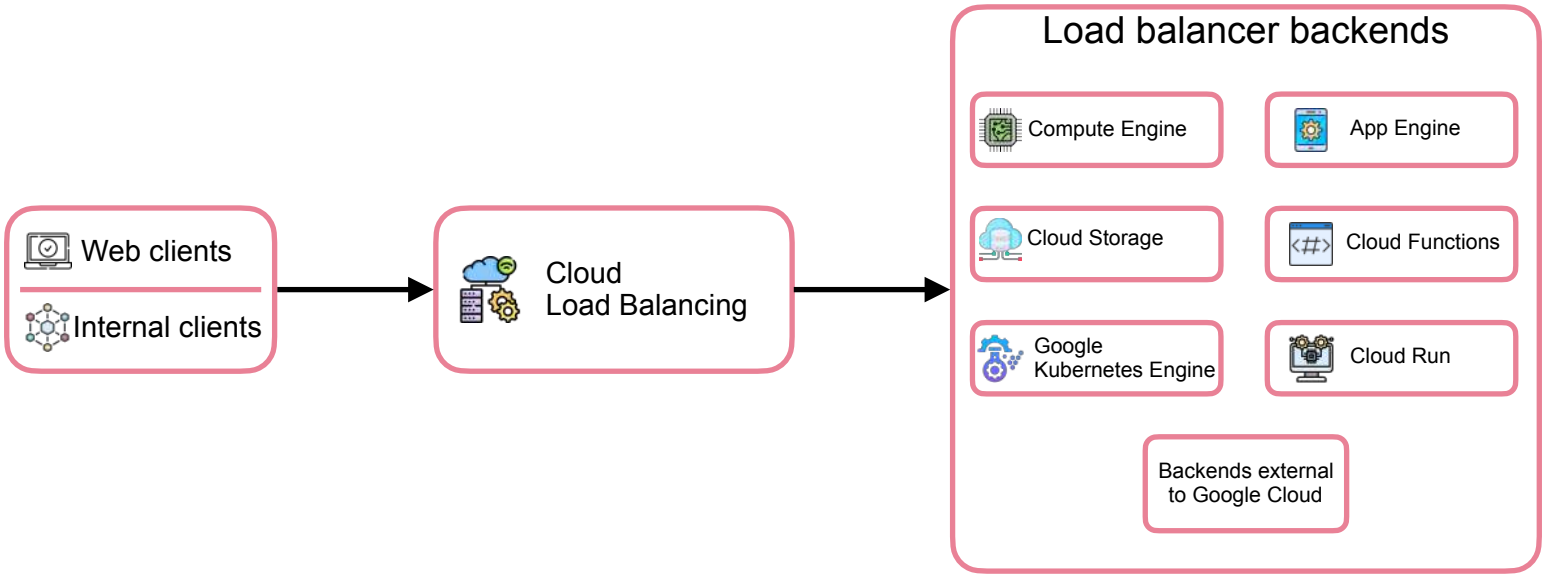
- Networking on the Google Cloud
- Interconnecting Networks
- Storage Solutions
- Identity and Access Management
  - IAM Best Practices
- Security Features Overview
- Logging and Monitoring


O'REILLY®
Load
Balancing

Load Balancers



Load Balancers Used with Multiple Backends



Load Balancers

- Complex service
- Many moving parts
- Basic idea
  - Stable front-end IP
  - Forwarding rules to funnel traffic
  - Connect to backend service
  - Distribute load intelligently
  - Health checks to avoid unhealthy instances

Load balancers **distribute** traffic to resources close to users and meet **high-availability** requirements

Load Balancers on the GCP

- Fully managed, software-defined, redundant and highly available
- Supports > 1 million queries per second with high performance and low latency
- Autoscaling to meet increased traffic

Load Balancing Categorization

Global

Use when your users and instances are globally distributed, load balanced resources lie across multiple regions

Regional

Use when instances and users are concentrated in one region

Load Balancing Categorization

External

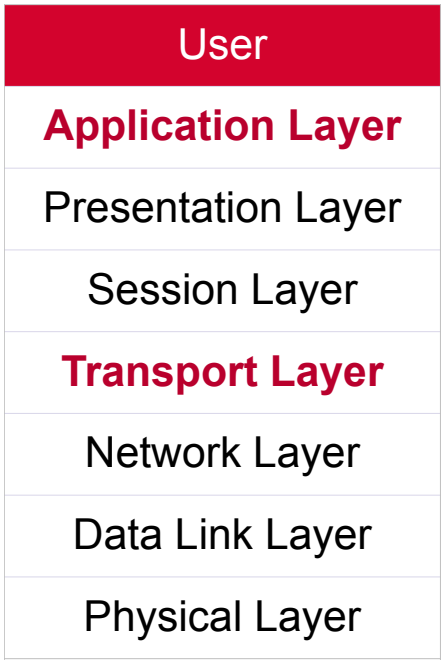
Distributes traffic from the internet to the Google Cloud

Internal

Distributes traffic only within the Google Cloud, all clients are inside of the Google Cloud

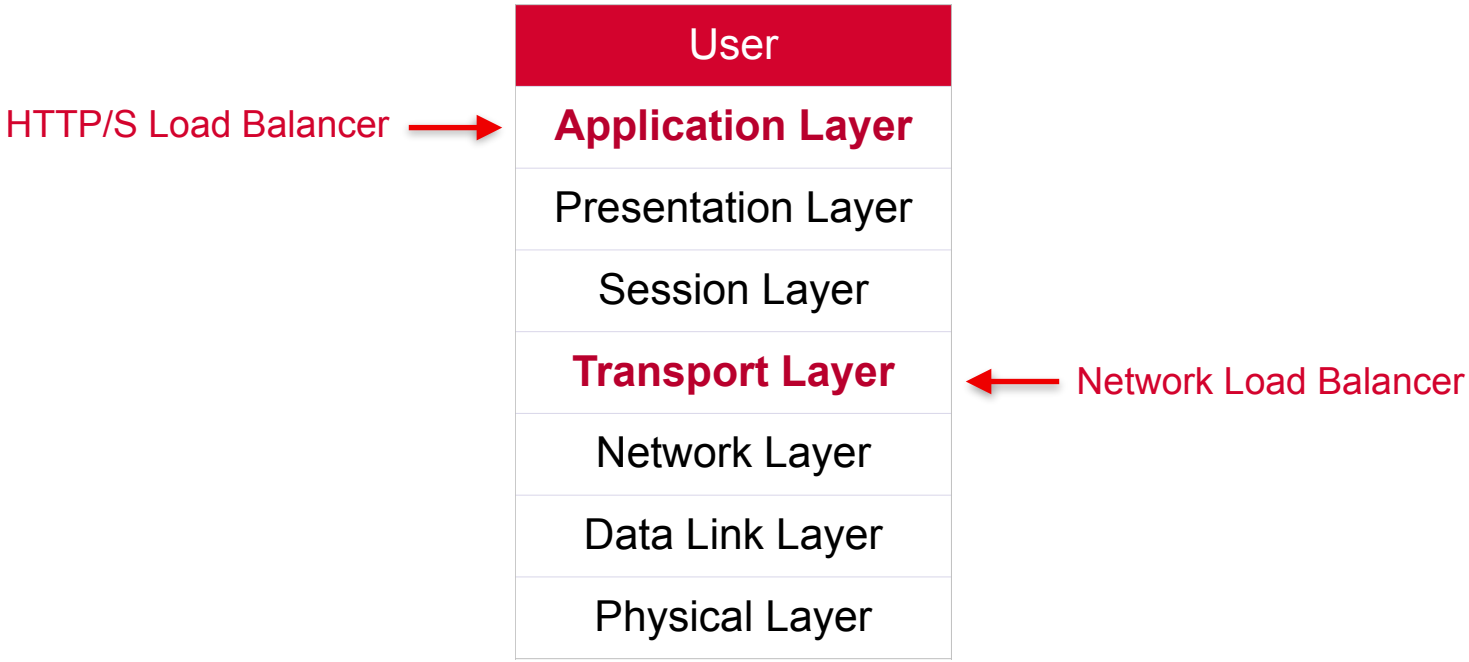
7 Layer OSI Network Stack

Routing decisions based on attributes of the request i.e. HTTP headers and the URL

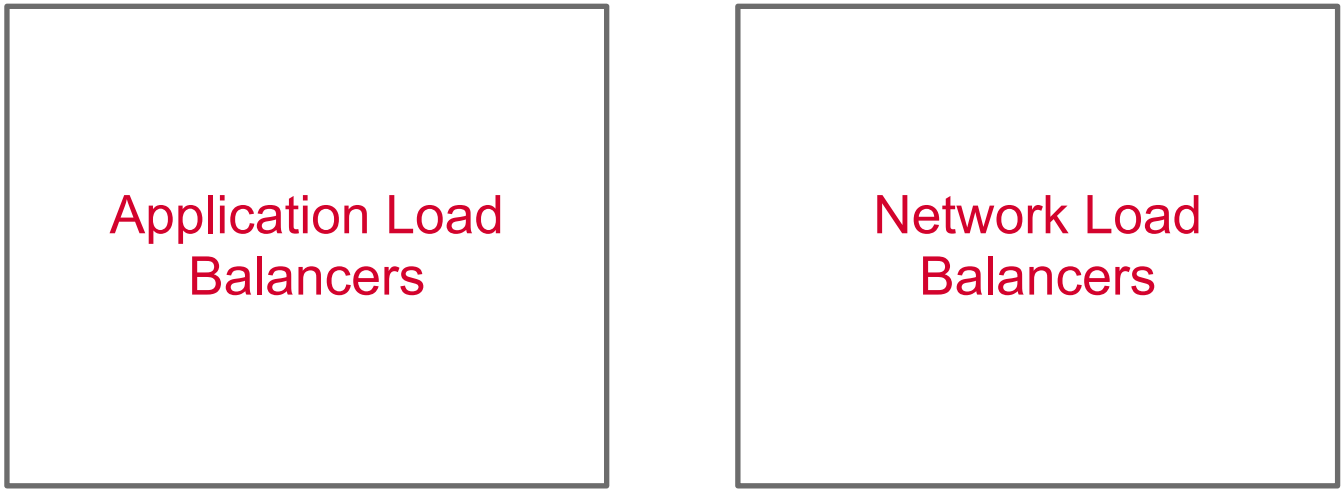


Direct traffic based on data from network and transport layer protocols such as TCP, UDP, ESP, GRE, ICMP, and ICMPv6

7 Layer OSI Network Stack



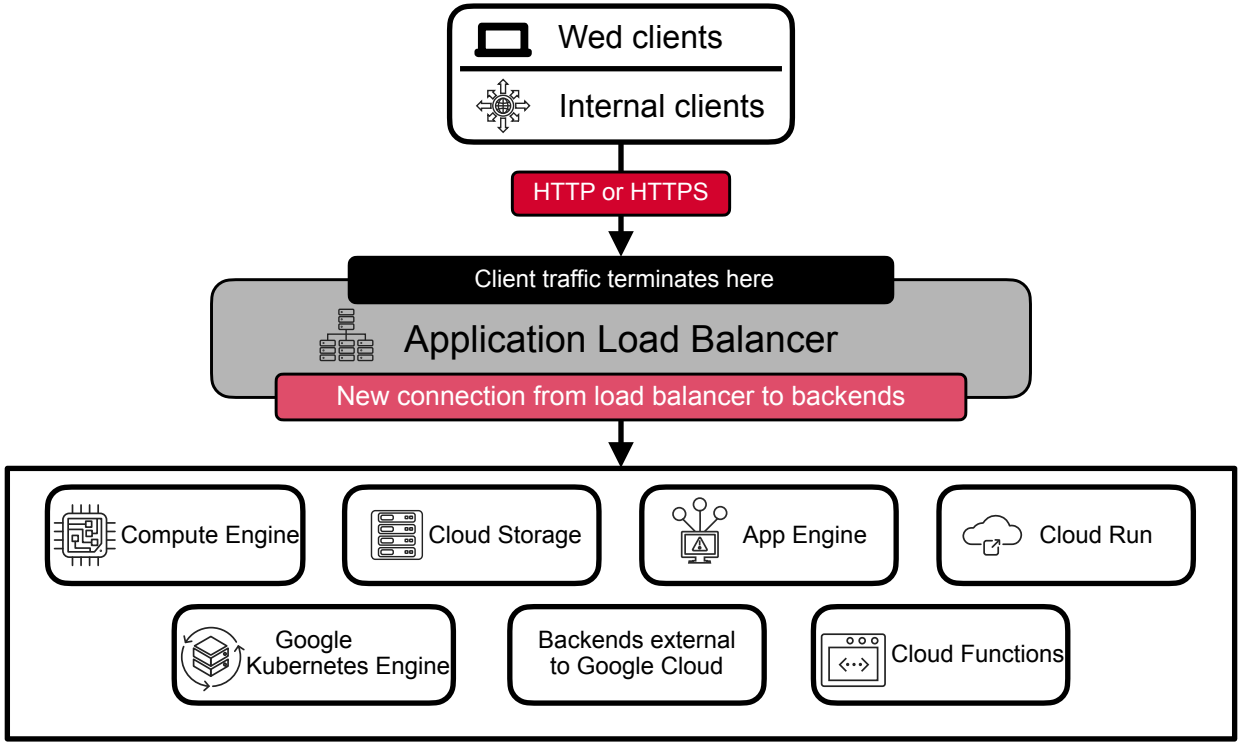
Two Types of Load Balancers



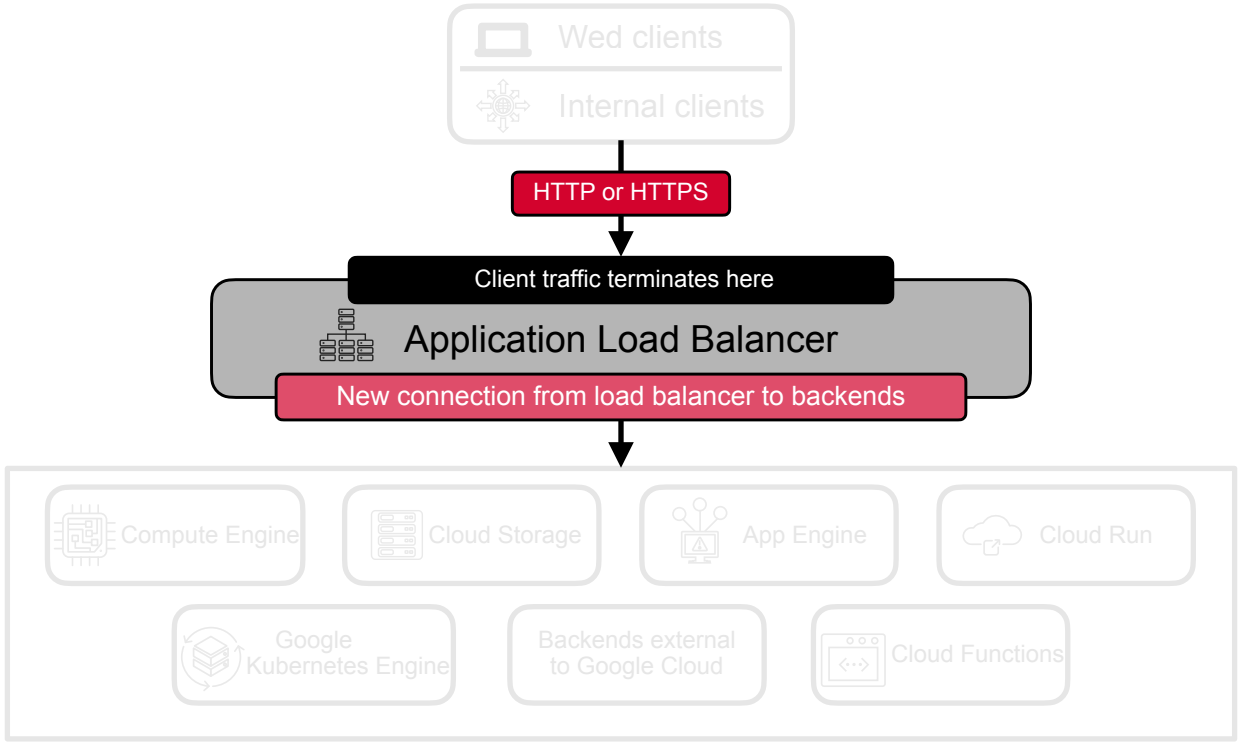
Application Load Balancers

- **Proxy-based** layer 7 load balancers
- Allow you to scale your services behind a single IP
- Distributes HTTP and HTTPS traffic to Google backends and external backends
  - Compute Engine, GKE, Cloud Run

Application Load Balancers



Proxy Load Balancing



URL-based Routing

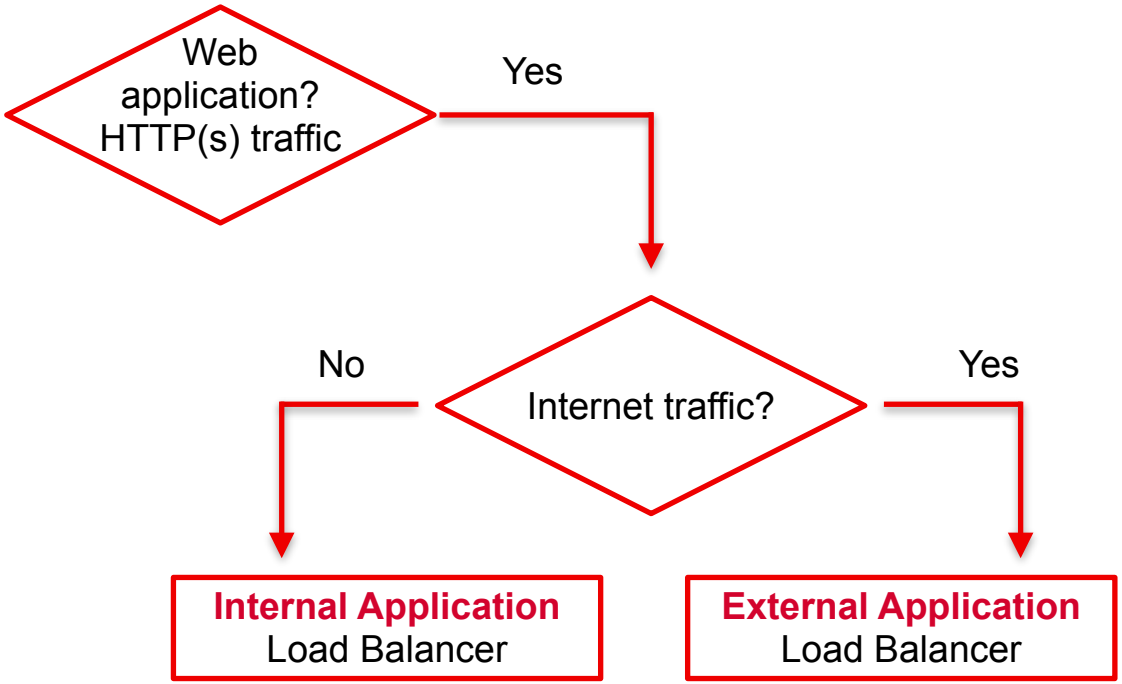
The **HTTP(S) Load Balancer** can split traffic based on content using **URL-based routing rules**

The load balancer inspects the incoming request's URL path or hostname and direct the traffic to different backend services or instances based on predefined conditions

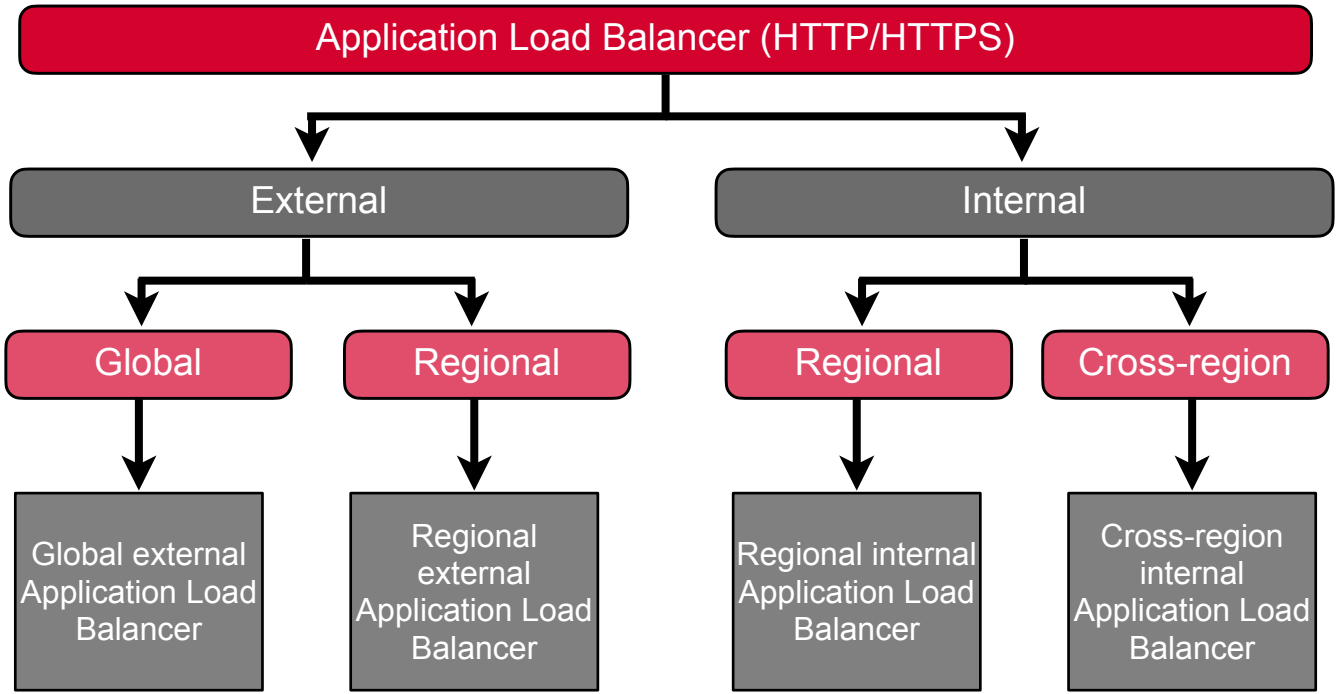
URL-based Routing

- Path-based routing
  - example.com/images/* routed to Backend Service 1
  - example.com/videos/* routed to Backend Service 2
- Host-based routing
  - app.example.com routed to Backend Service 3
  - blog.example.com routed to Backend Service 4

Choosing Load Balancers



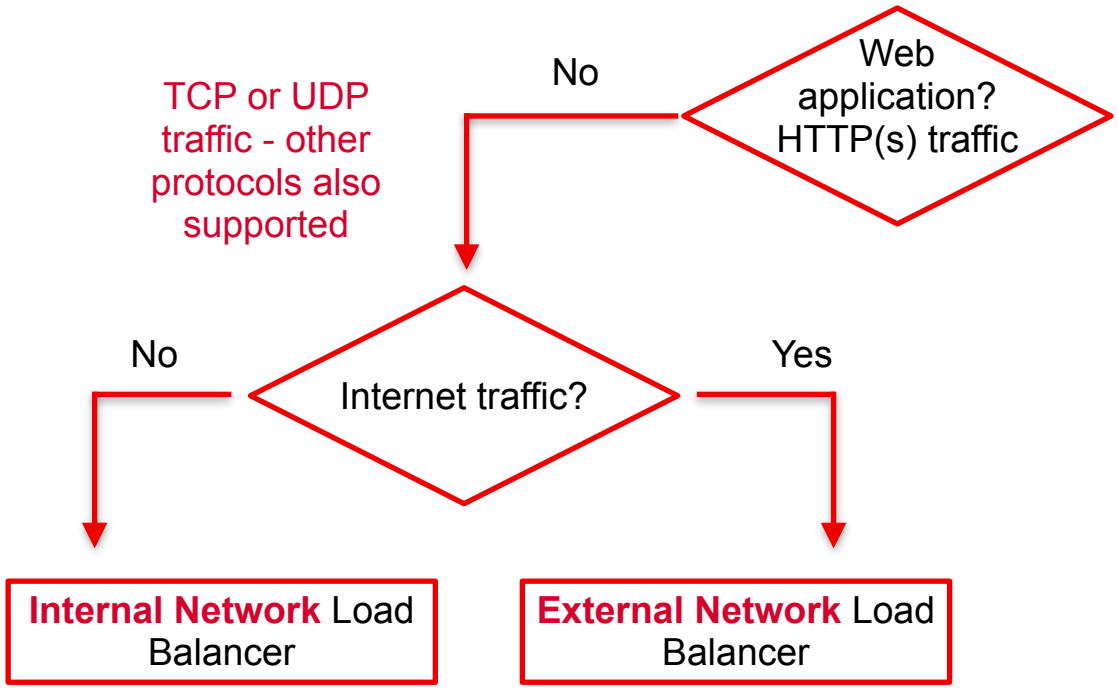
Application Load Balancers



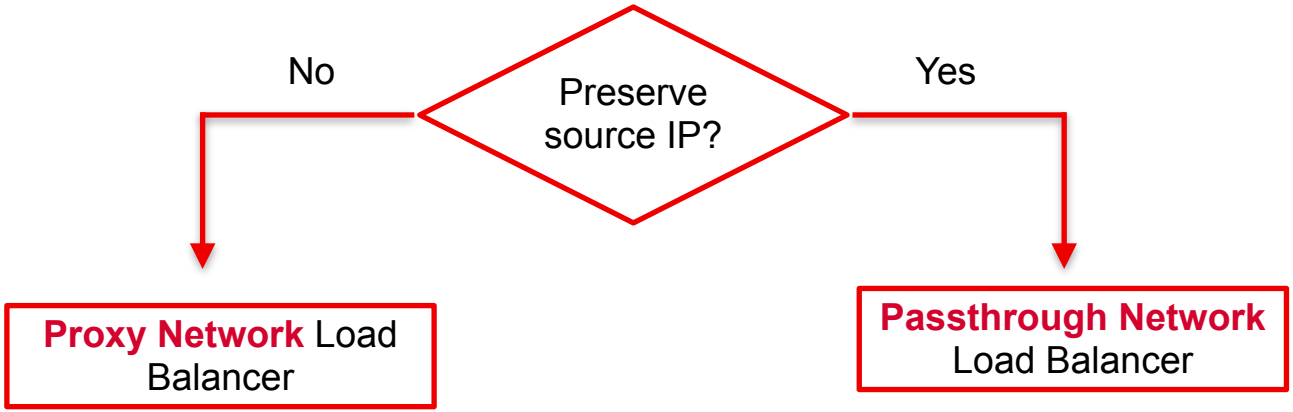
Network Load Balancers

- Layer 4 load balancers
- Handle TCP, UDP, or other IP protocol traffic
- Can be of two types
    - Proxy
    - Passthrough

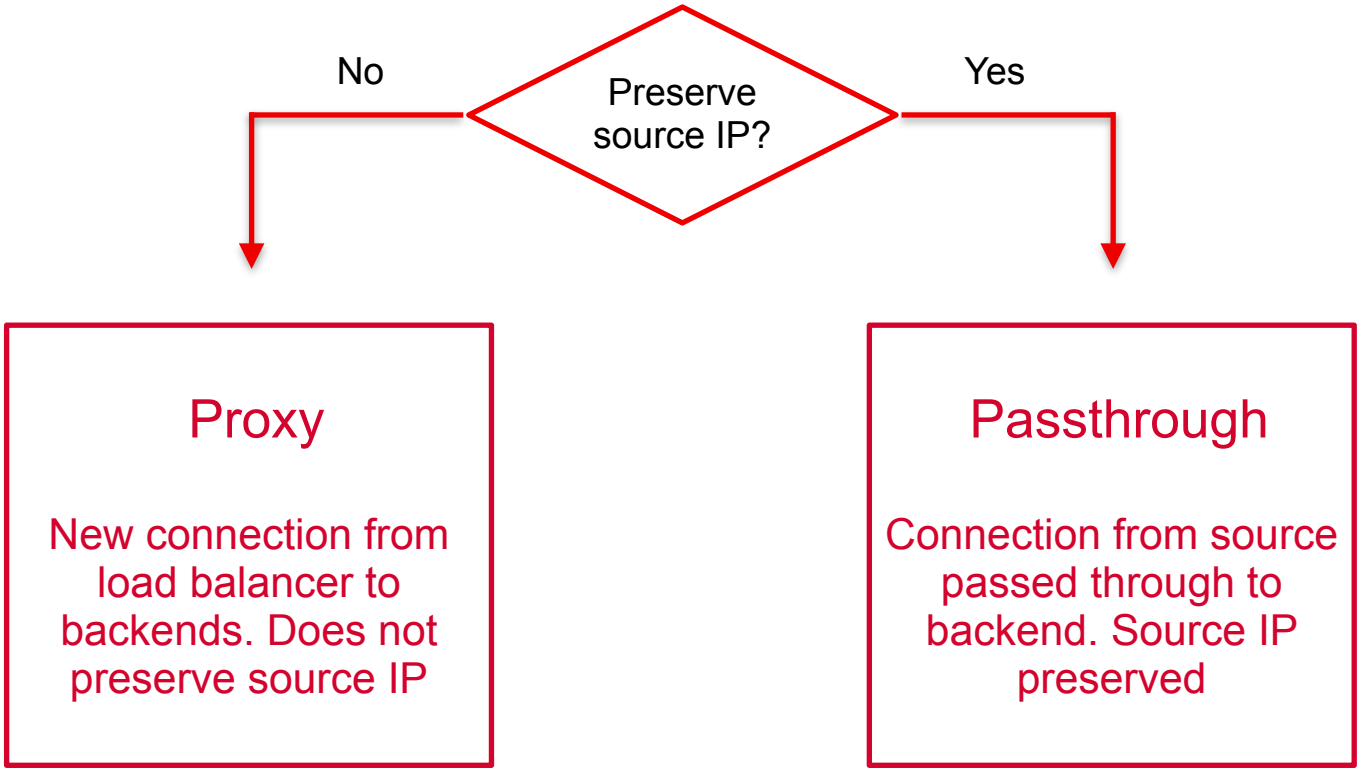
Choosing Load Balancers



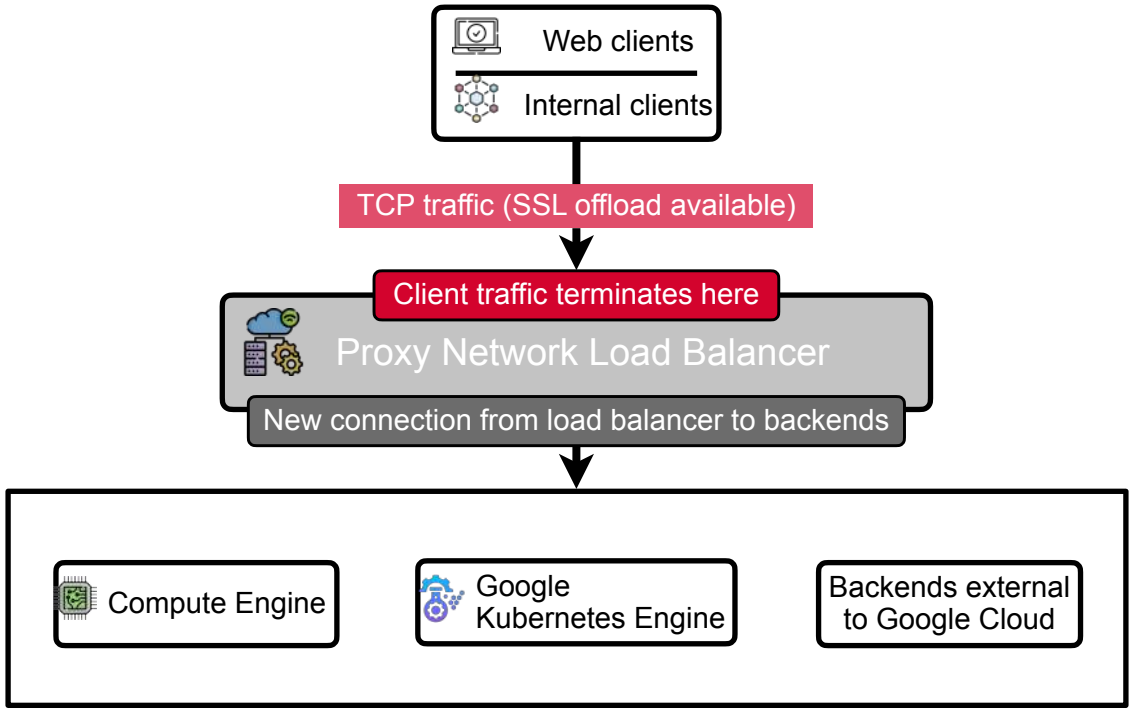
Choosing Network Load Balancers



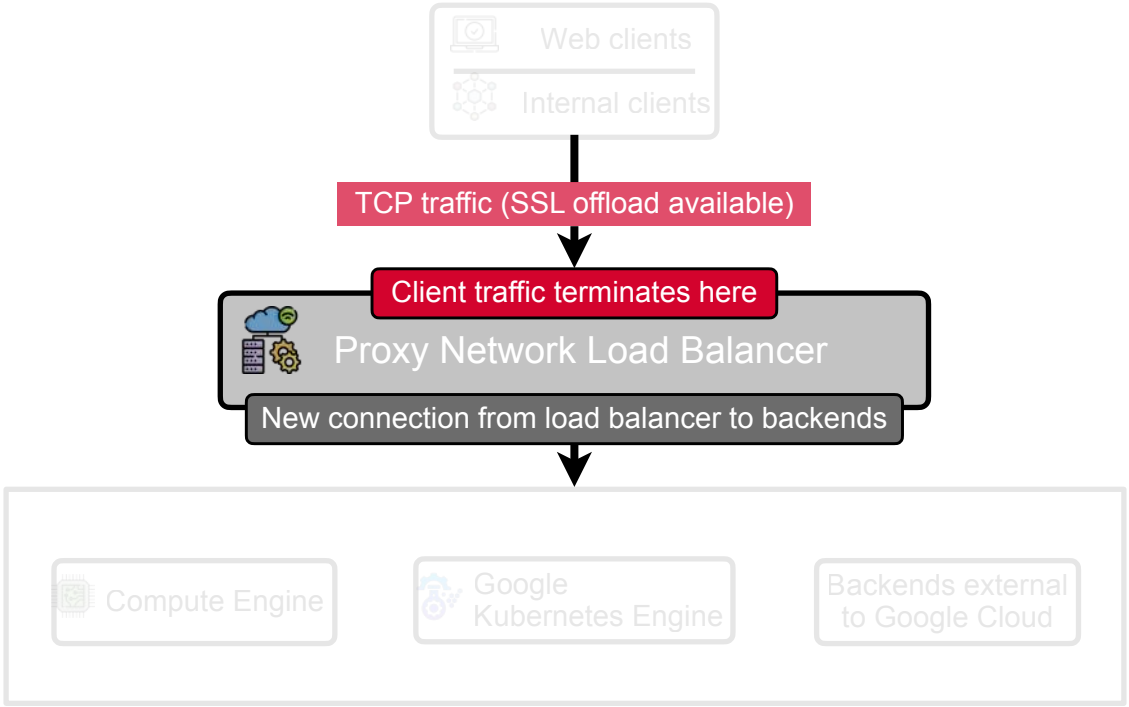
Choosing Network Load Balancers



Network Proxy Load Balancers

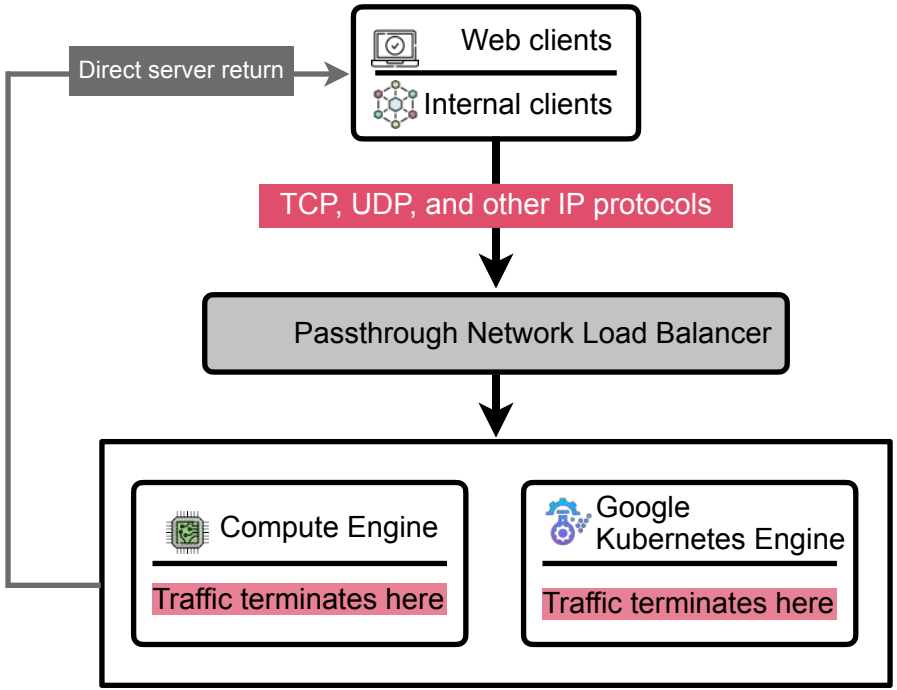


Network Proxy Load Balancers - SSL Offload

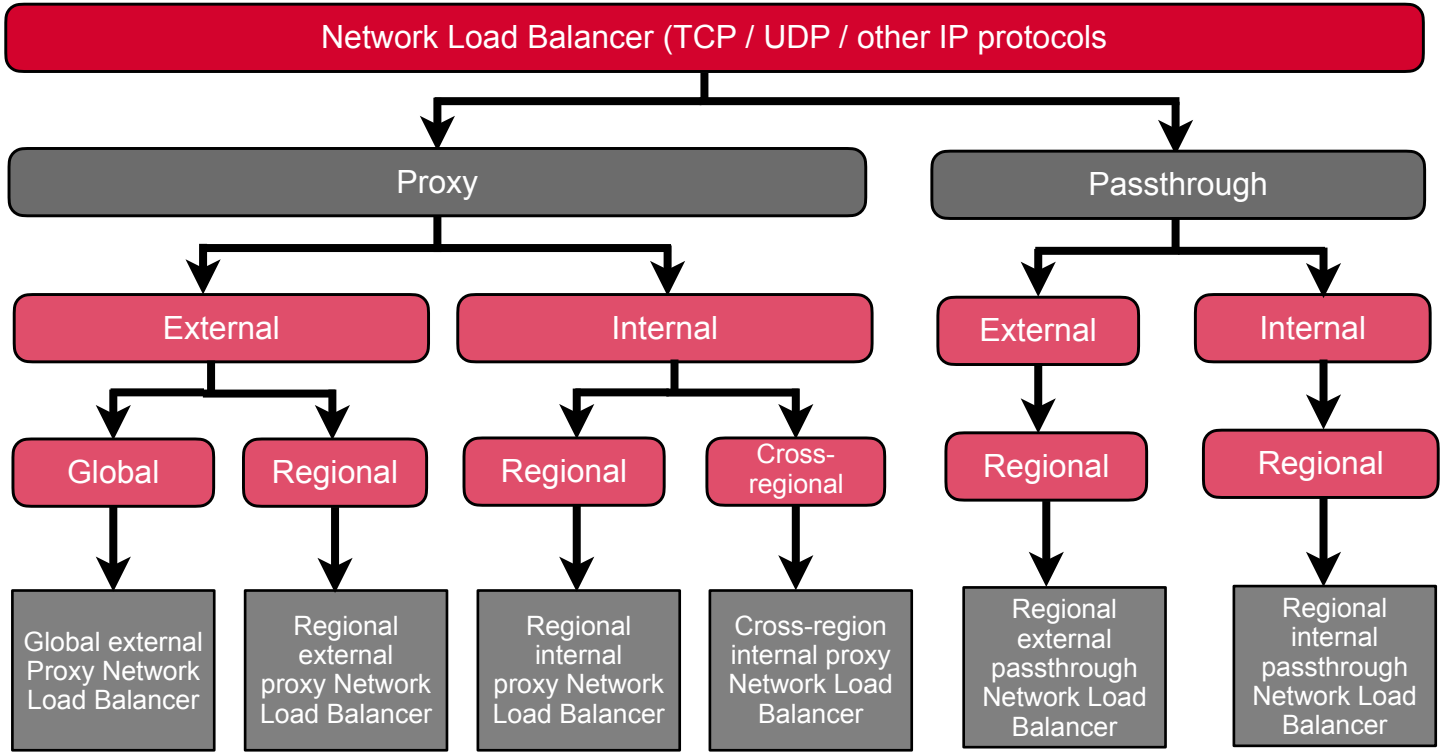


The load balancer terminates the secure SSL/TLS connection, decrypts the data, and forwards the plain HTTP traffic to the backend servers.

Network Passthrough Load Balancers



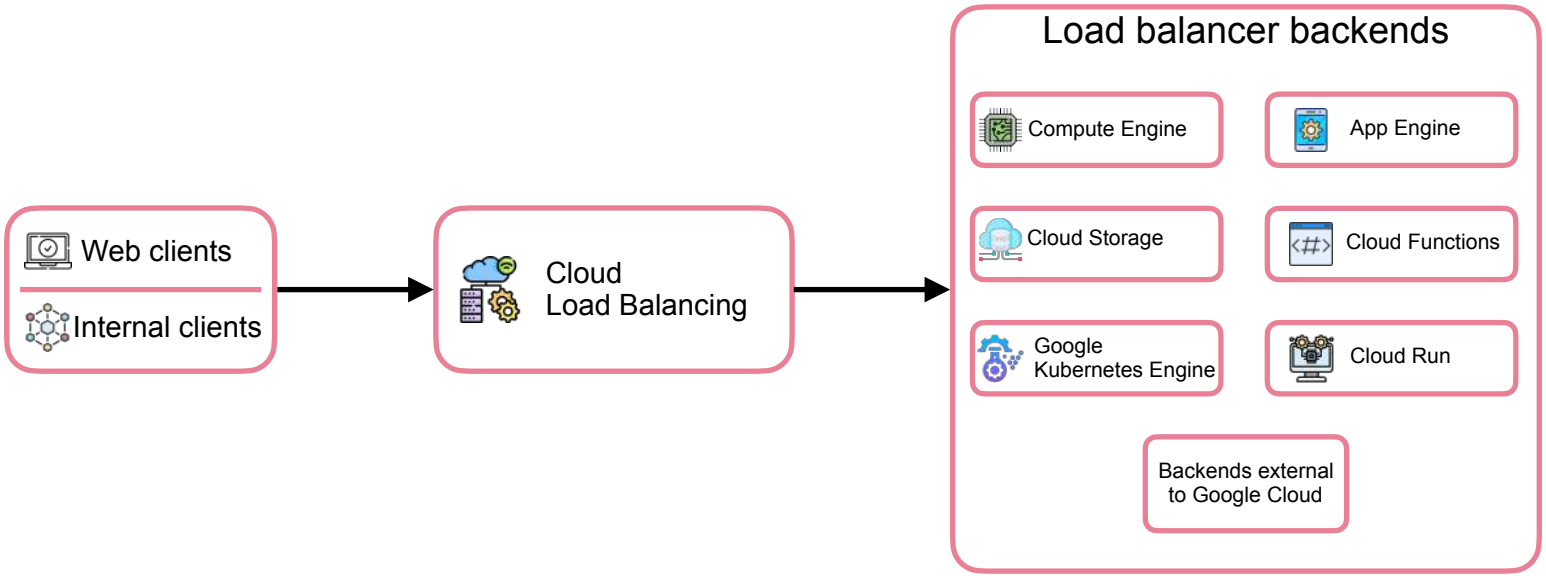
Network Load Balancers



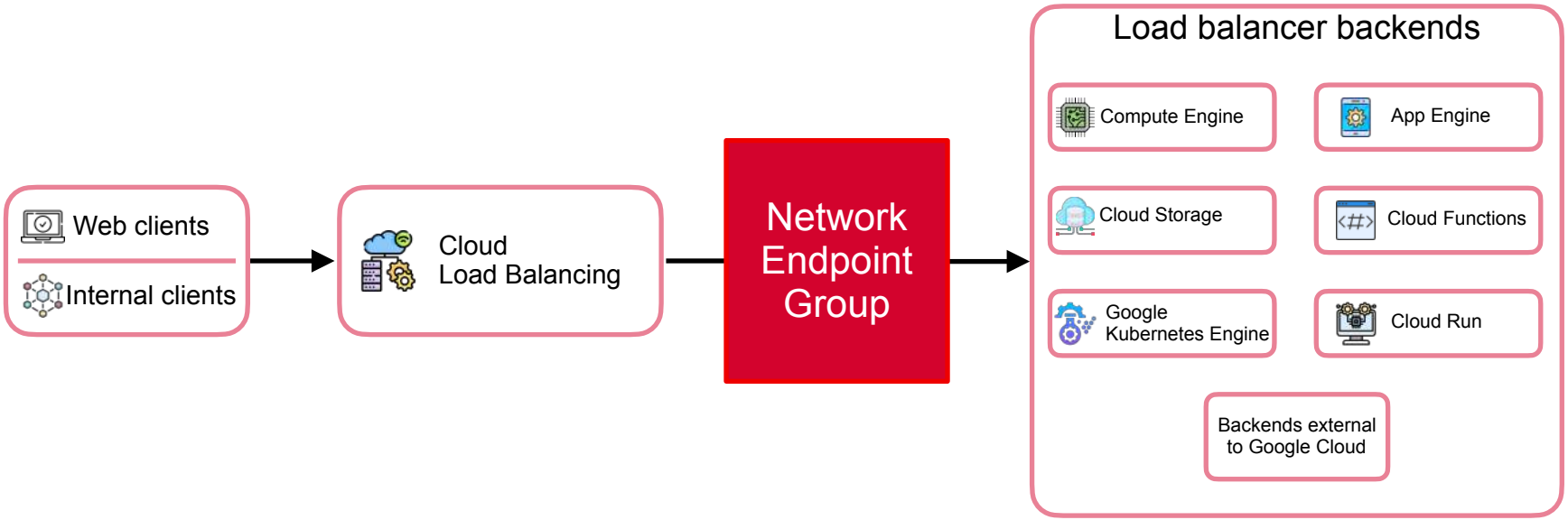
Network Endpoint Groups

Logical grouping of network endpoints, where an endpoint is an `IP_address:Port` combination. They provide a flexible and scalable way to group backend endpoints for your load balancers.

Load Balancers Used with Multiple Backends

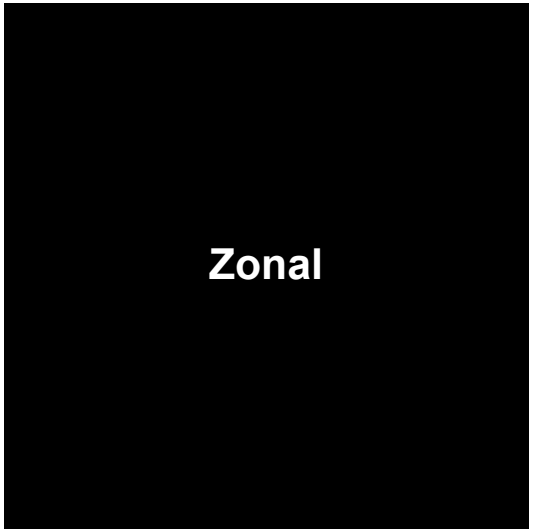


Decouple Load Balancer from Backend



Types of NEGs

Types of NEGs

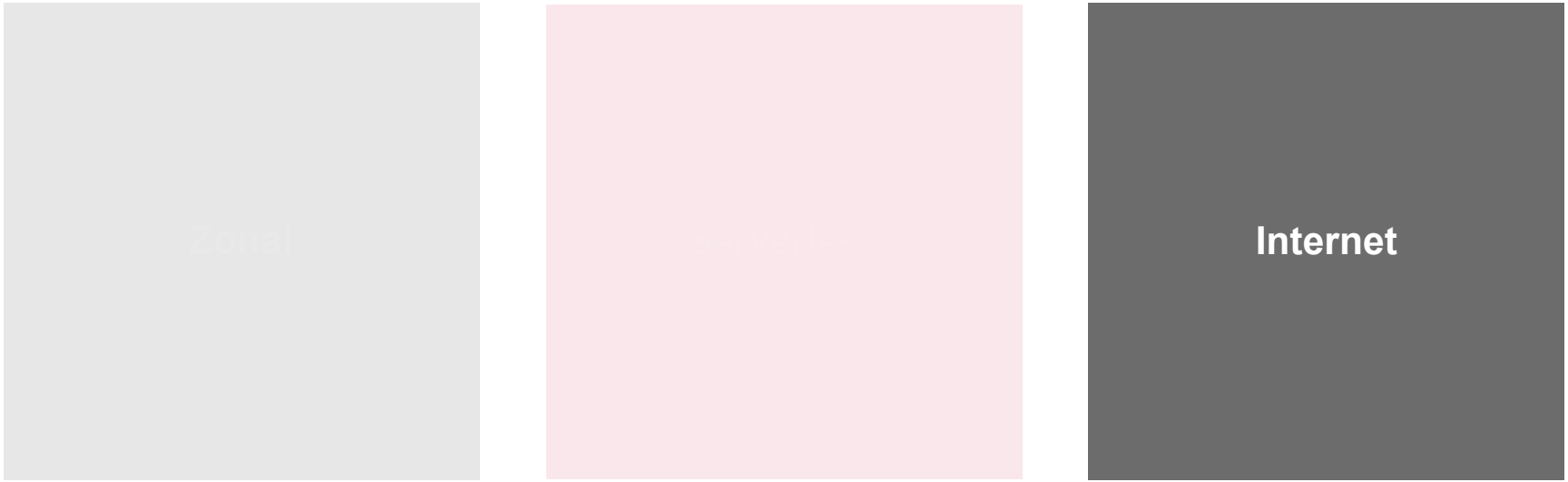


A group of endpoints that are located within a single zone. Used for container-native load balancing on zonal GKE clusters

Types of NEGs

Points to a single serverless Google Cloud service, such as Cloud Run, App Engine, or Cloud Functions.

Types of NEGs



A group of endpoints that live **outside of Google Cloud**, on the public internet. You can specify endpoints by either their IP address or Fully Qualified Domain Name (FQDN).

When you configure the different types of Cloud Load Balancers for your services the right type of NEG is automatically created and used

Load Balancing

Your team needs to deploy a legacy financial application on Google Cloud. The application communicates over TCP and relies on direct access to a local filesystem to maintain data integrity. It cannot be scaled horizontally because it does not have synchronization and accessing the file system concurrently causes problems that cannot be resolved easily. Although brief downtime is acceptable during failover, the application must remain highly available to support continuous business operations. How should you architect this deployment?

A. Deploy the application on a managed instance group across multiple zones, use Cloud Filestore for shared storage, and place an HTTP(S) load balancer in front to distribute traffic.

B. Deploy the application on a managed instance group across multiple zones, use Cloud Filestore for shared storage, and use a network load balancer to balance traffic.

C. Create an unmanaged instance group with one active and one standby VM in different zones, attach a regional persistent disk for storage, and place an HTTP(S) load balancer in front to route requests.

D. Create an unmanaged instance group with an active VM and a standby VM in separate zones, use a regional persistent disk for data storage, and place a network load balancer in front to route client connections.

Load Balancing

Your team needs to deploy a legacy financial application on Google Cloud. The application communicates over TCP and relies on direct access to a local filesystem to maintain data integrity. It cannot be scaled horizontally because it does not have synchronization and accessing the file system concurrently causes problems that cannot be resolved easily. Although brief downtime is acceptable during failover, the application must remain highly available to support continuous business operations. How should you architect this deployment?

A. Deploy the application on a managed instance group across multiple zones, use Cloud Filestore for shared storage, and place an HTTP(S) load balancer in front to distribute traffic.

B. Deploy the application on a managed instance group across multiple zones, use Cloud Filestore for shared storage, and use a network load balancer to balance traffic.

C. Create an unmanaged instance group with one active and one standby VM in different zones, attach a regional persistent disk for storage, and place an HTTP(S) load balancer in front to route requests.

D. Create an unmanaged instance group with an active VM and a standby VM in separate zones, use a regional persistent disk for data storage, and place a network load balancer in front to route client connections.

Load Balancing

A media company is building a global video streaming platform on Google Cloud using dozens of microservices. They require a CI/CD pipeline for frequent, independent updates using immutable artifacts. The architecture must provide low-latency access to users in both North America and Asia via a single global entry point, while keeping core backend services private from the internet. Which set of Google Cloud services is most suitable for this architecture?

A. Artifact Registry to store container images, Google Kubernetes Engine (GKE) clusters in an Asian and North American region, and a Global External HTTP/S Load Balancer.

B. App Engine to host the services in multiple regions, Cloud Spanner for a global database, and a regional Internal TCP/UDP Load Balancer.

C. Cloud Storage to store application binaries, Managed Instance Groups (MIGs) in each region, and a Regional External Network Load Balancer for each region.

D. Cloud Functions for the microservice logic and a Global External HTTP/S Load Balancer configured with Serverless Network Endpoint Groups (NEGs).

Load Balancing

A media company is building a global video streaming platform on Google Cloud using dozens of microservices. They require a CI/CD pipeline for frequent, independent updates using immutable artifacts. The architecture must provide low-latency access to users in both North America and Asia via a single global entry point, while keeping core backend services private from the internet. Which set of Google Cloud services is most suitable for this architecture?

A. **Artifact Registry to store container images, Google Kubernetes Engine (GKE) clusters in an Asian and North American region, and a Global External HTTP/S Load Balancer.**

B. App Engine to host the services in multiple regions, Cloud Spanner for a global database, and a regional Internal TCP/UDP Load Balancer.

C. Cloud Storage to store application binaries, Managed Instance Groups (MIGs) in each region, and a Regional External Network Load Balancer for each region.

D. Cloud Functions for the microservice logic and a Global External HTTP/S Load Balancer configured with Serverless Network Endpoint Groups (NEGs).

Load Balancing

An international e-commerce company has deployed its containerized shopping cart API on Cloud Run in two regions: europe-west1 and australia-southeast1. They need to provide a single global endpoint (api.shopping.com) for their frontend applications that automatically routes customers to the closest healthy region, ensuring a fast and resilient shopping experience.

What is the recommended Google Cloud native approach to achieve this?

A. Create a **Regional External HTTP/S Load Balancer** in both regions. Use Cloud DNS to create weighted A records to distribute traffic between the two load balancer IPs.

B. For each regional Cloud Run service, create a **Serverless Network Endpoint Group (NEG)**. Create a single **Global External HTTP/S Load Balancer** and configure its backend service to use these two Serverless NEGs.

C. Use **API Gateway** in front of each Cloud Run service. Configure the two API Gateway instances under a single custom domain in Cloud DNS to handle routing.

D. Create a **Global External HTTP/S Load Balancer**. Manually add the default .run.app URLs of the two Cloud Run services as **Internet Network Endpoint Group** backends to the load balancer.

Load Balancing

An international e-commerce company has deployed its containerized shopping cart API on Cloud Run in two regions: europe-west1 and australia-southeast1. They need to provide a single global endpoint (api.shopping.com) for their frontend applications that automatically routes customers to the closest healthy region, ensuring a fast and resilient shopping experience.

What is the recommended Google Cloud native approach to achieve this?

A. Create a **Regional External HTTP/S Load Balancer** in both regions. Use Cloud DNS to create weighted A records to distribute traffic between the two load balancer IPs.

B. For each regional Cloud Run service, create a **Serverless Network Endpoint Group (NEG)**. Create a single **Global External HTTP/S Load Balancer** and configure its backend service to use these two Serverless NEGs.

C. Use **API Gateway** in front of each Cloud Run service. Configure the two API Gateway instances under a single custom domain in Cloud DNS to handle routing.

D. Create a **Global External HTTP/S Load Balancer**. Manually add the default .run.app URLs of the two Cloud Run services as **Internet Network Endpoint Group** backends to the load balancer.

O'REILLY®

Networking

IP addresses, routes, and firewall rules all exist inside a GCP resource called a VPC Network

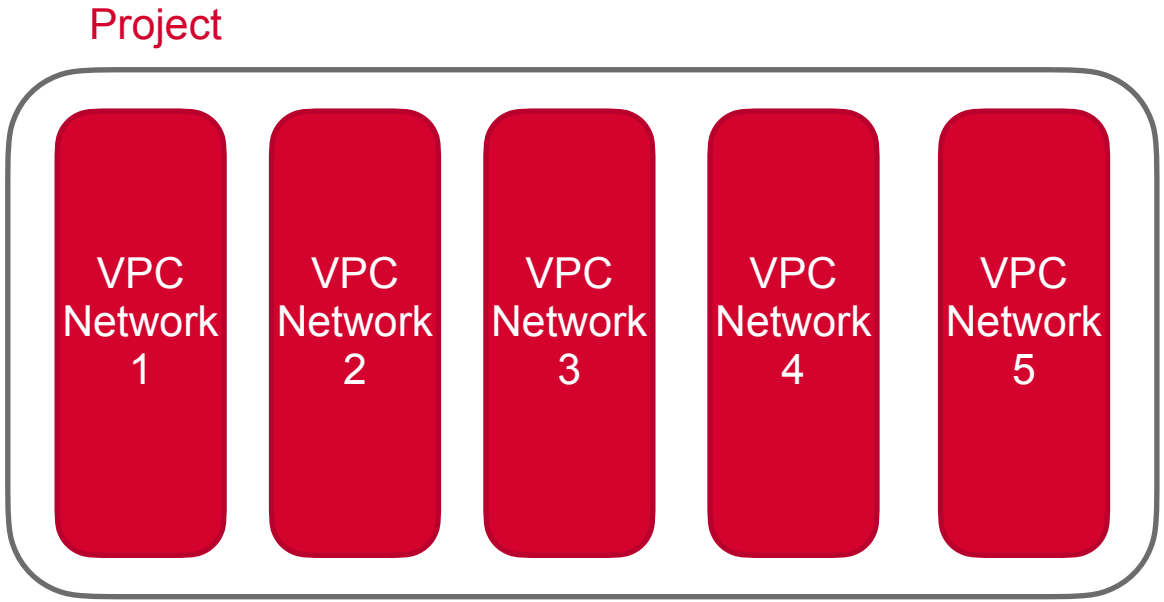
Google Virtual Private Cloud

A VPC network, often just called a network, is a global, private, isolated virtual network partition that provides managed network functionality

Google Virtual Private Cloud

A VPC network, often just called a network, is a global, private, isolated virtual network partition that provides managed network functionality

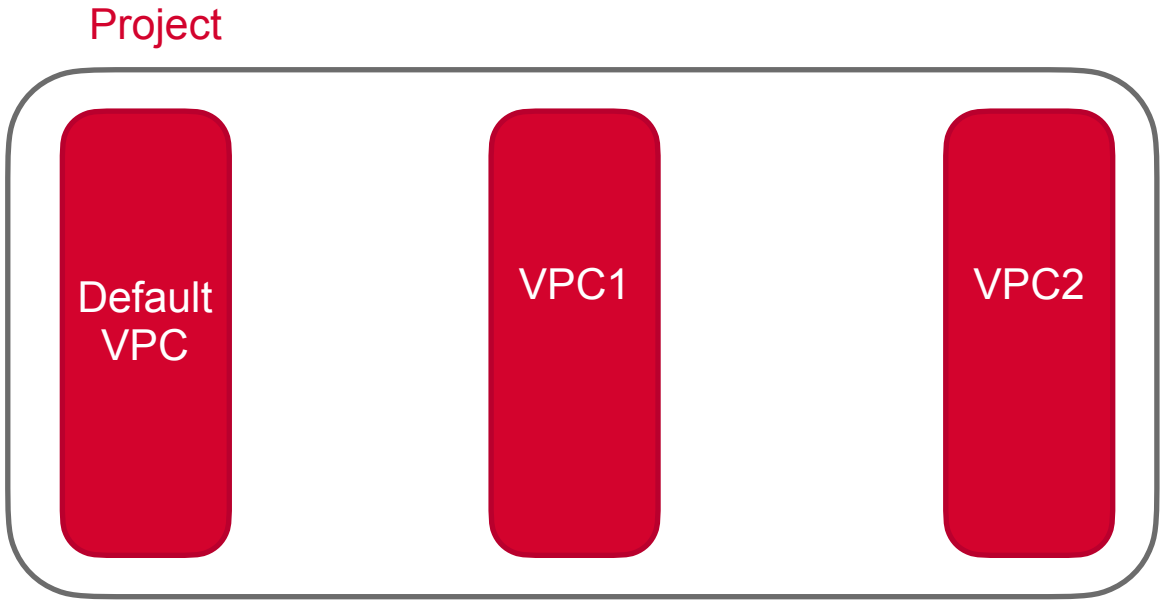
Multiple VPCs in a Project



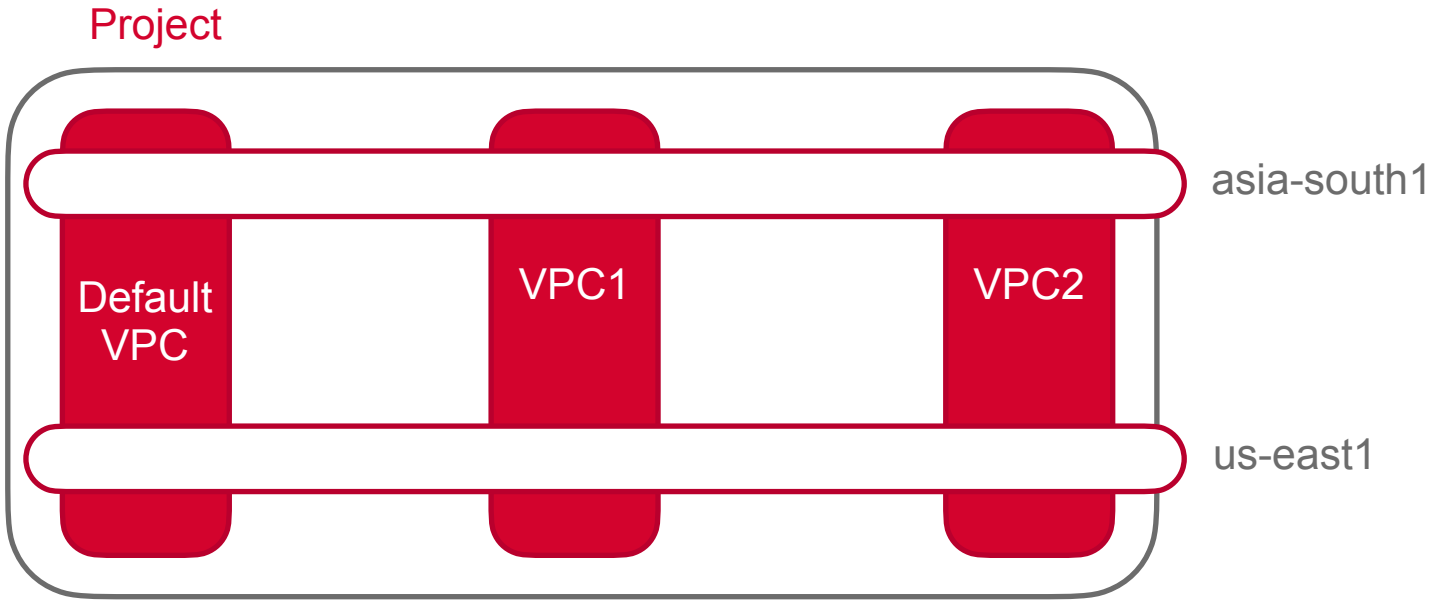
Projects and VPCs

- VPCs are global resources on the GCP
- Each VPC must exist inside a project
- Default VPC pre-created in each project

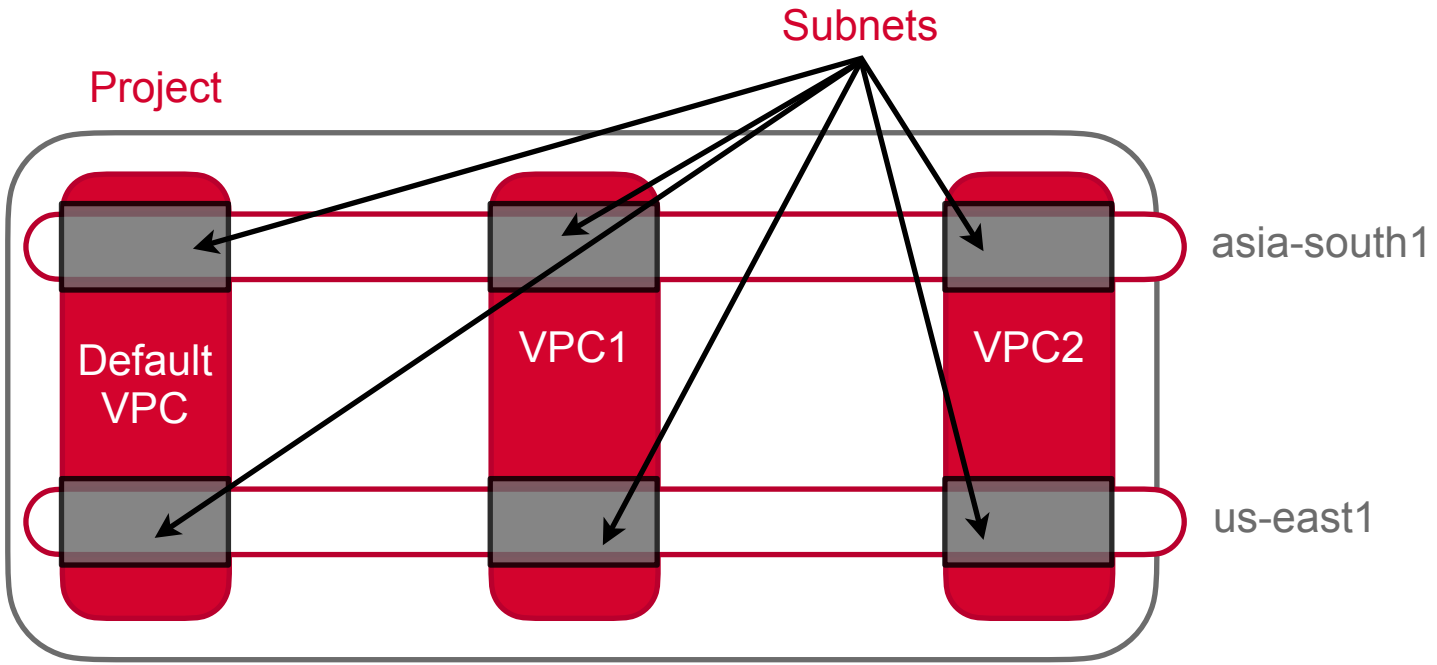
VPCs Are Global



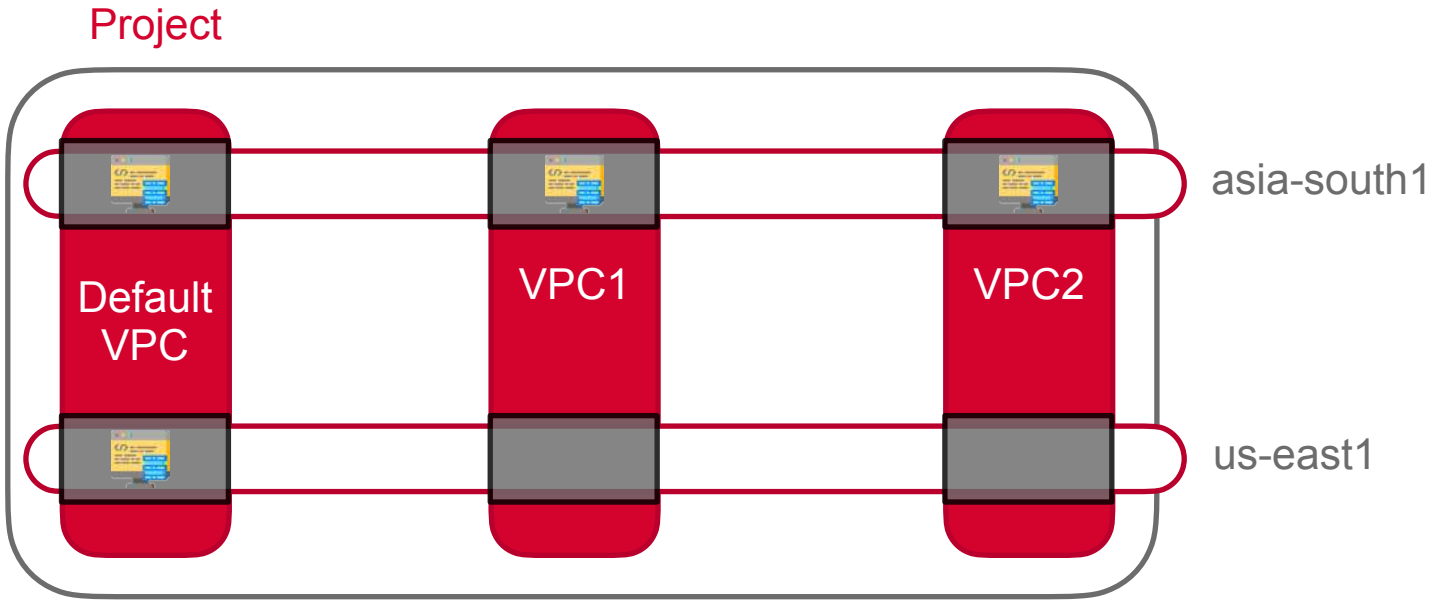
VPCs Are Global



Subnets in Each Region



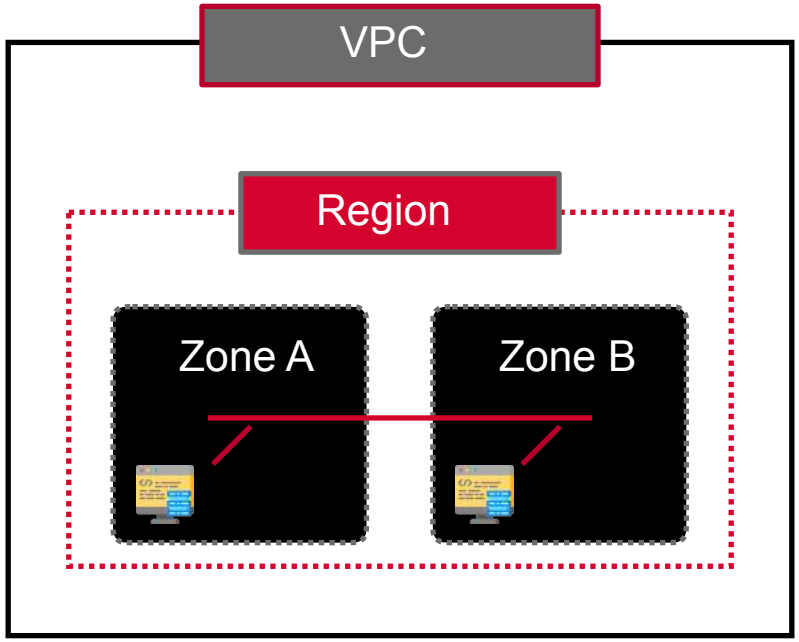
Resources Provisioned on Subnets



Subnets

- **IP range partitions** within global VPCs
- VPCs have no IP ranges
- Subnets are regional - can span zones inside a region
- Network has to have at least one subnet before you can use it

Subnets Span Zones



Subnets and IP Ranges

- Each subnet must have primary address range
- Valid RFC 1918 CIDR block
- Subnet ranges in same network cannot overlap
- Subnet ranges in different networks can overlap

AutoMode and CustomMode VPCs

Auto Mode

Subnets automatically created in each region, default firewall rules

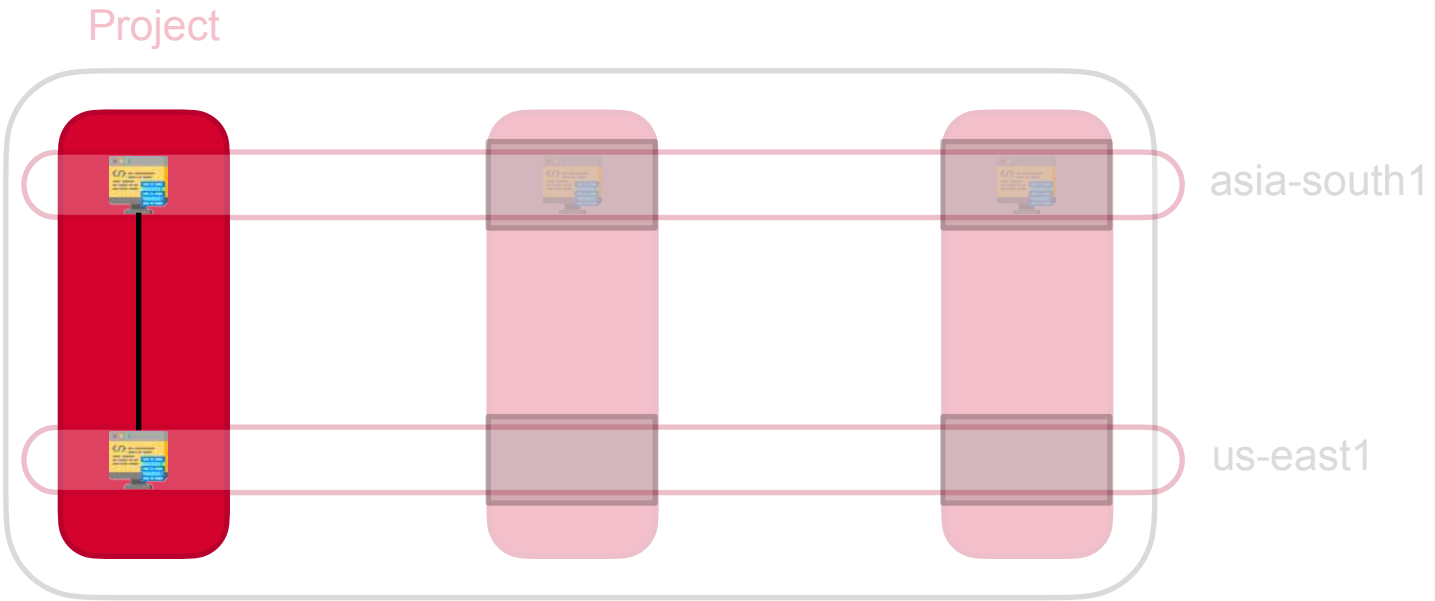
Custom Mode

Manually create subnets in regions, no defaults preconfigured

Auto Mode and Custom Mode VPCs

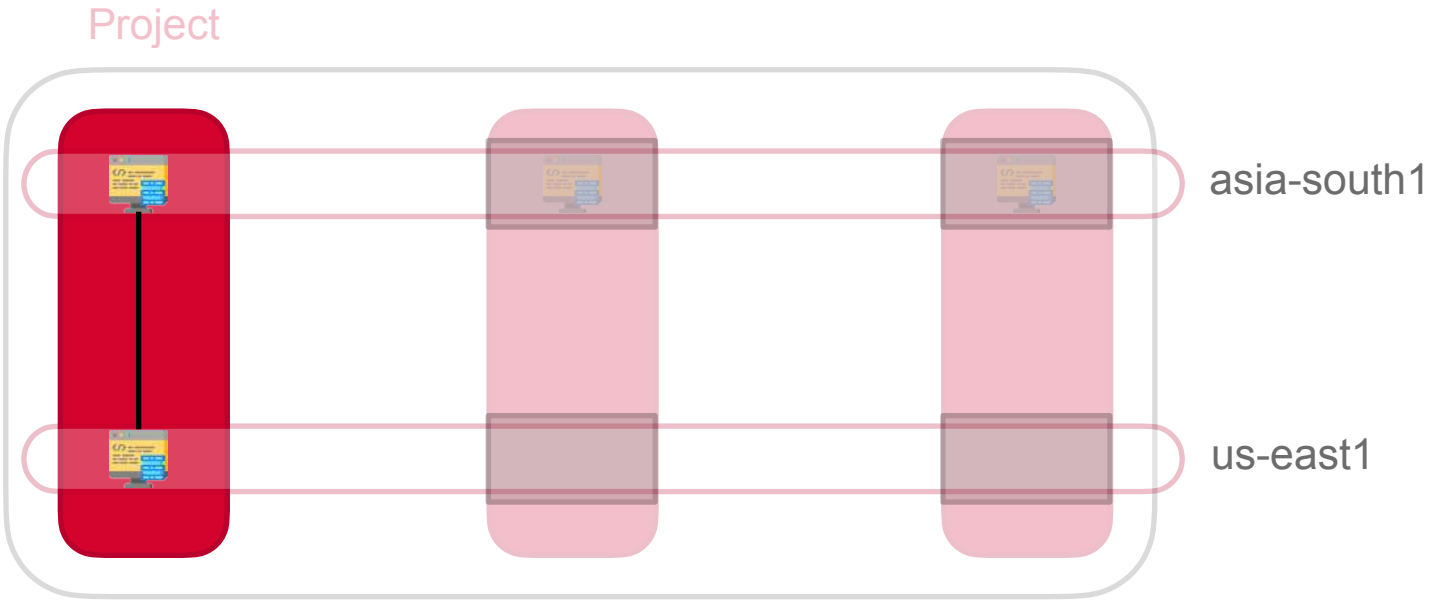
- Auto Mode VPCs have pre-created subnets
  - One in each GCP region
- Custom Mode VPCs start with no subnets
  - Full control over which regions have subnets
  - Can create multiple subnets in a region

Communication on VPCs



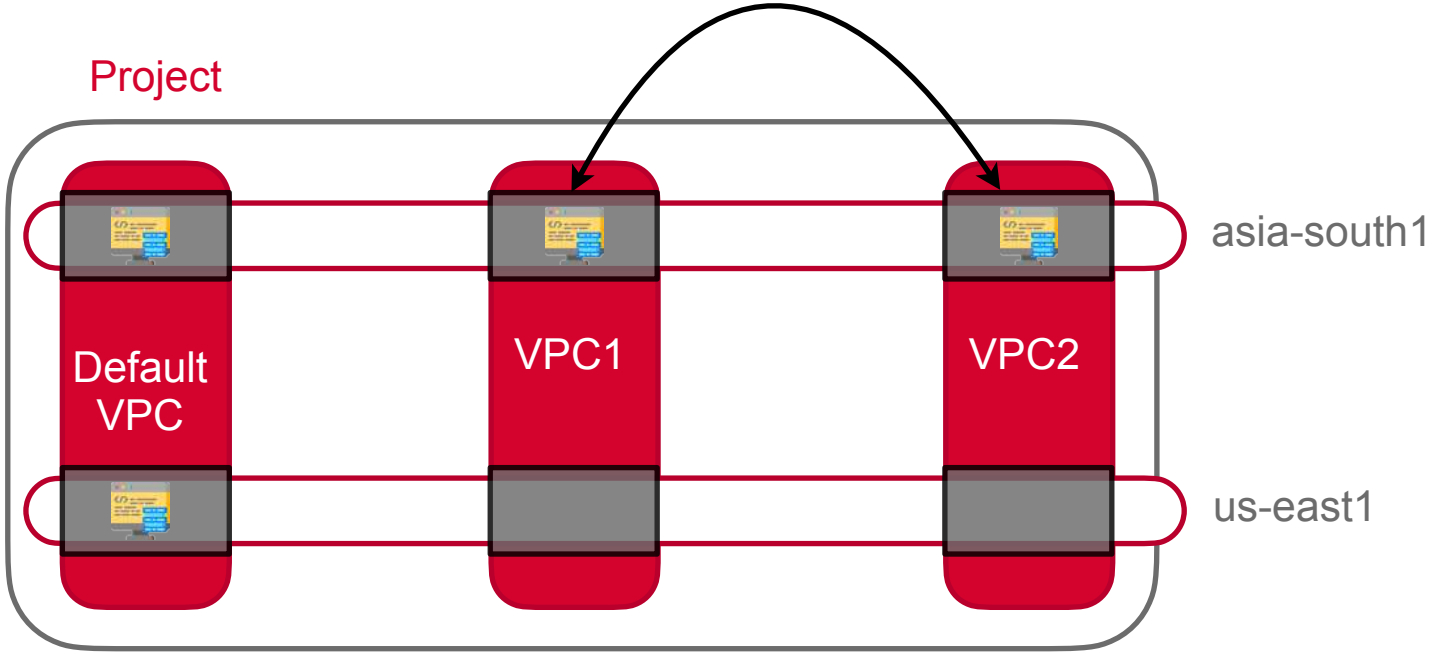
Resources within a VPC communicate using private IP addresses

Communication on VPCs



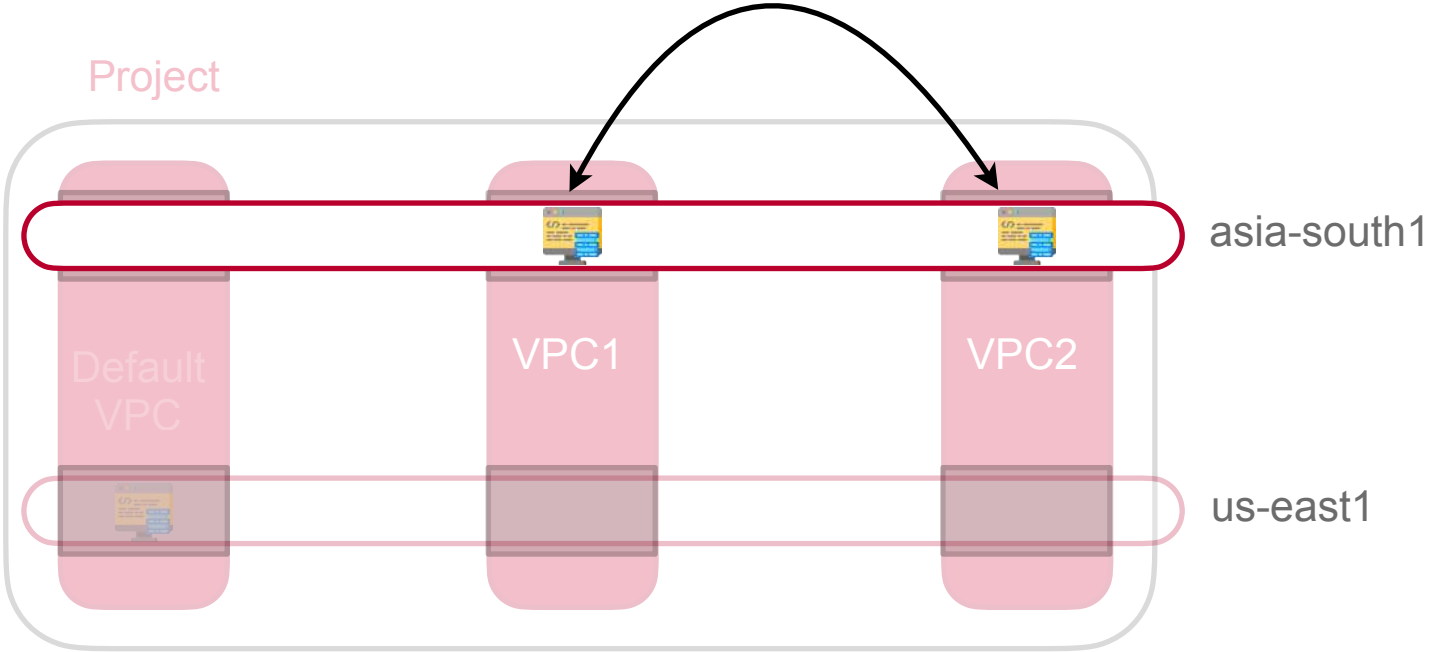
Wherever they are located in the world -
irrespective of physical location

Communication on VPCs



Resources on different VPCs communicate over the internet using external IPs

Communication on VPCs



Even though they are in the same region - they may even be in the same zone on the same physical hardware

Default VPC

- Pre-created on every project
- Includes subnet for each GCP region
- New subnets added when new regions are created
- Resources created here by default

Default VPC

- Includes routes for all resources
- All VMs on the default VPC can talk to each other
- Default gateway to internet
- Includes several firewall rules

Firewall Rules

- Every VPC is a distributed firewall
- Firewall rules defined in VPC
- Are applied on per-instance basis
- Can also regulate internal traffic

Firewall Rules

- Every VPC has two permanent rules
  - Implied allow egress
  - Implied deny ingress
- Can be overridden by more specific rules
- In addition, default VPC has several rules

Additional Rules in Default VPC

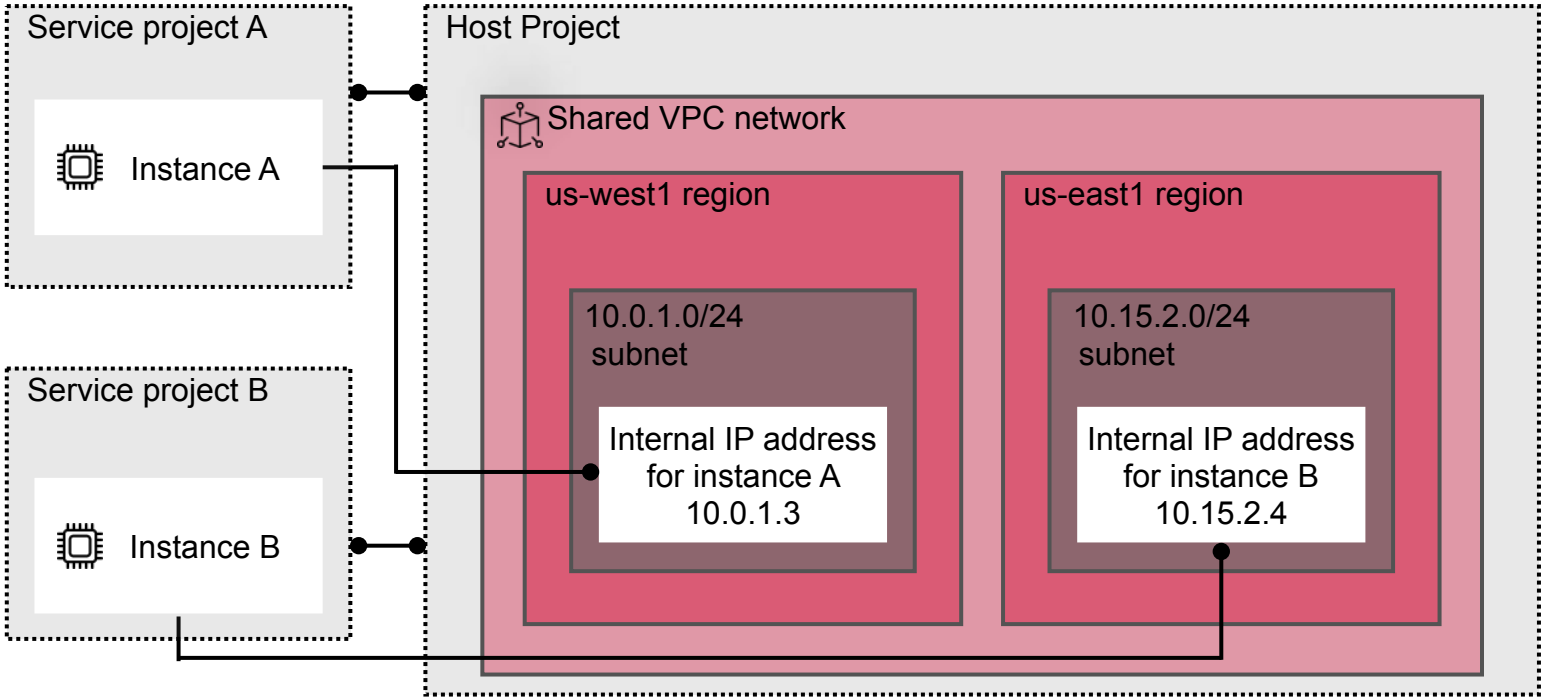
- default-allow-internal
- default-allow-ssh
- default-allow-rdp
- default-allow-icmp


O'REILLY®
Connecting Networks

Shared VPC

- Share VPC across projects on GCP
- One VPC shared across projects
- Projects must be in the same organization
- Host project, guest resources
- Shared VPC admin to administer the shared VPC

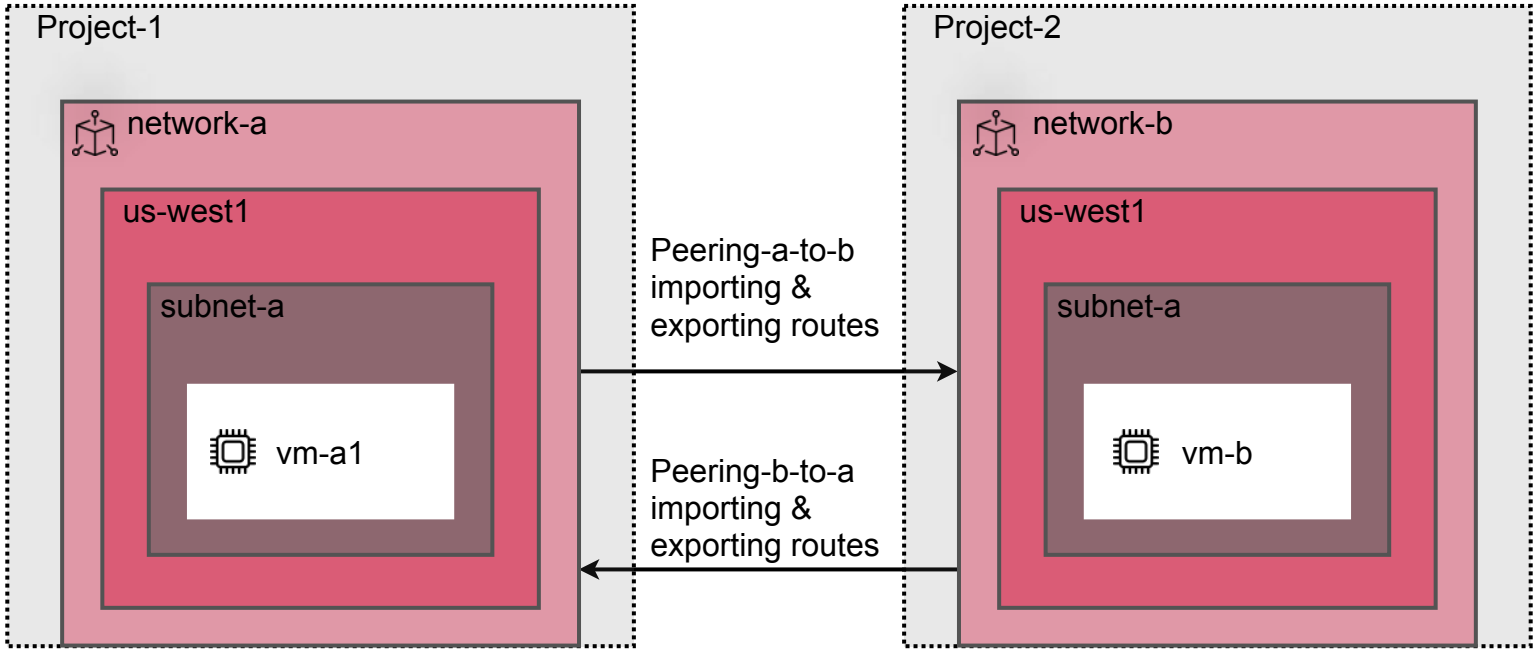
Shared VPC


Service project A
Instance A
Service project B
Instance B
Host Project
Shared VPC network
us-west1 region
10.0.1.0/24
subnet
Internal IP address
for instance A
10.0.1.3
us-east1 region
10.15.2.0/24
subnet
Internal IP address
for instance B
10.15.2.4

VPC Peering

- Two or more VPCs shared across projects
- Projects need not be in the same organization
- Allows resources on different VPC networks to communicate using internal IP addresses
- Resources on the network use Google infrastructure to communicate
- Reduced latency, higher security and lower cost as compared with using external IPs

VPC Peering


Project-1
network-a
us-west1
subnet-a
vm-a1
Peering-a-to-b
importing &
exporting routes
Peering-b-to-a
importing &
exporting routes
Project-2
network-b
us-west1
subnet-a
vm-b

Shared VPCs vs. Network Peering

Shared VPCs

- Only within **same organization**
- One VPC used across projects
- Host and service projects not peers
- Single level of sharing possible

Network Peering

- Across **organization boundaries**
- Multiple VPCs share resources
- Connected VPCs are peers
- Multiple levels of peering possible

Shared VPC and VPC Peering both allow instances on the networks to communicate using internal IPs

Interconnecting Networks

GCP-to-GCP

VPC Network Peering

Enterprise connectivity

Peering and interconnect options

Interconnecting Networks

GCP-to-GCP

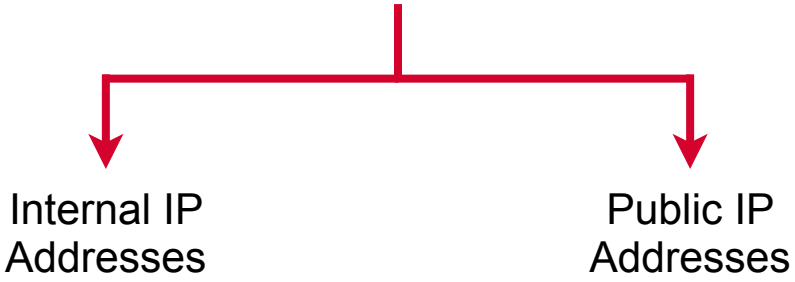
VPC Network Peering

Enterprise connectivity

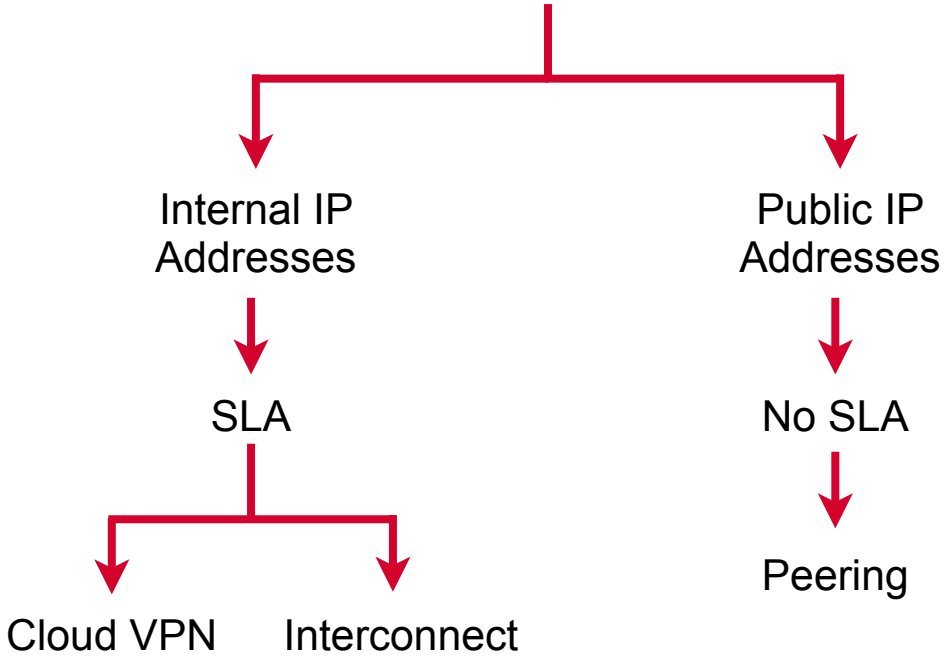
Peering and interconnect options

Connect a cloud network with an on-premise network using private or public IP addresses

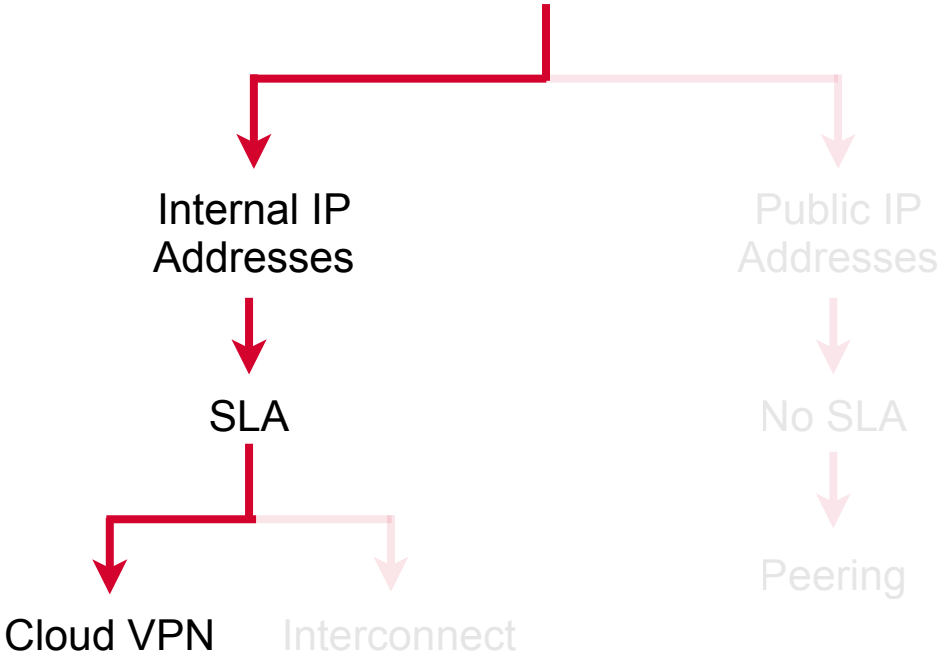
Enterprise Connectivity



Enterprise Connectivity



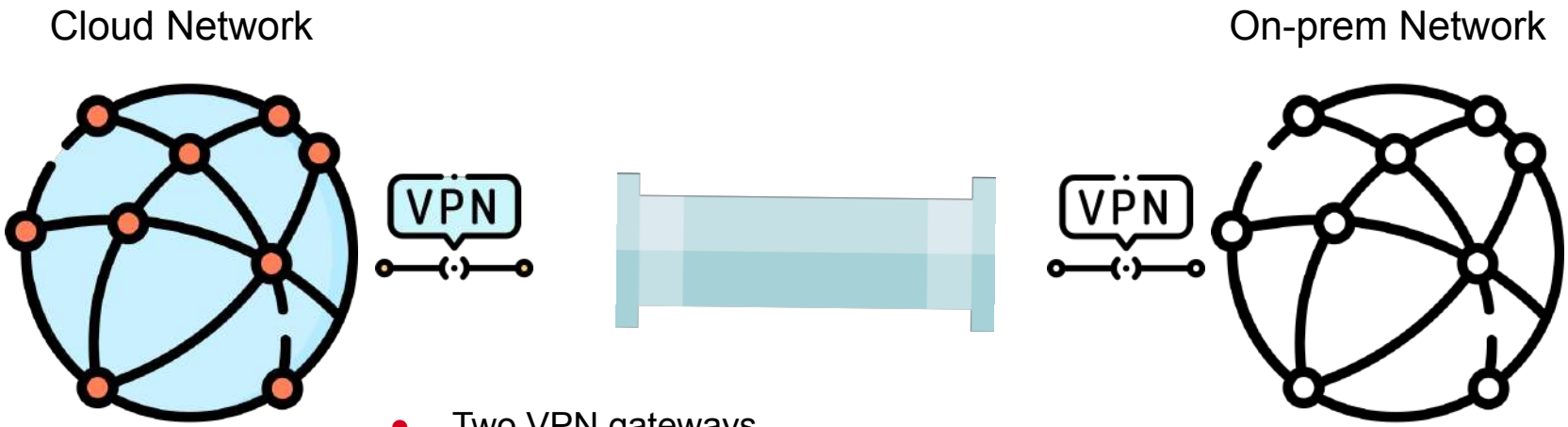
Enterprise Connectivity



Cloud VPN

Configuration Property	Choice
Connection	**Encrypted tunnel to VPC networks through the public internet**
Access Type	**Internal IP** addresses in RFC 1918 address space
Capacity	1.5-3 Gbps for each tunnel
Other Considerations	**Requires a VPN device on your on-premises network**

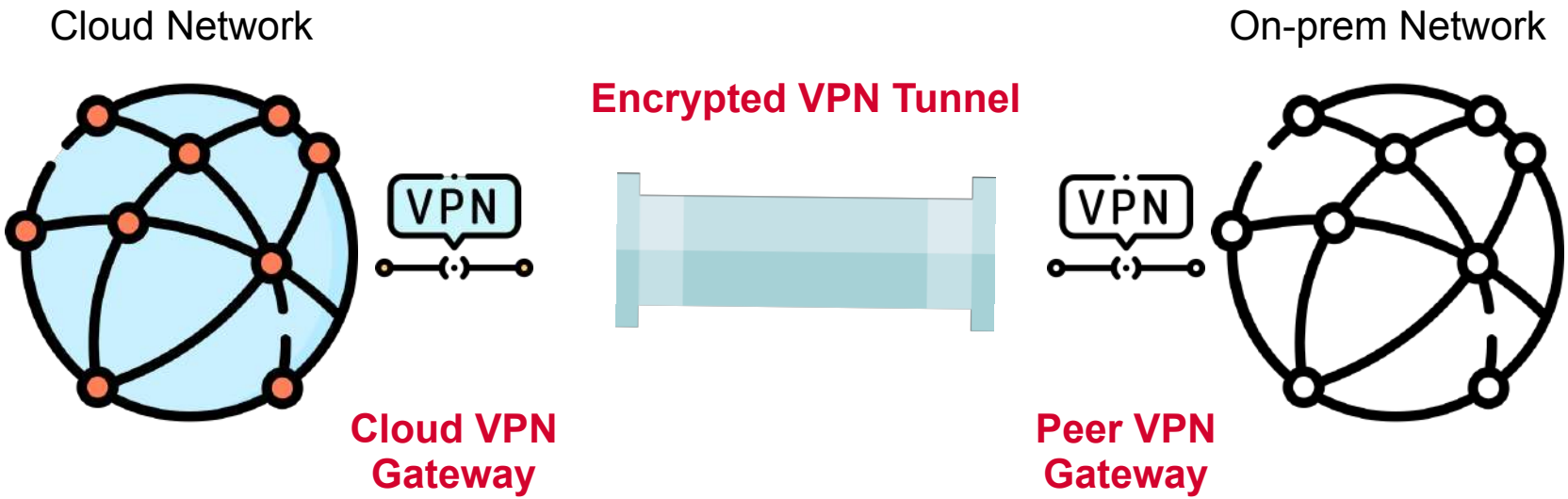
Cloud VPN



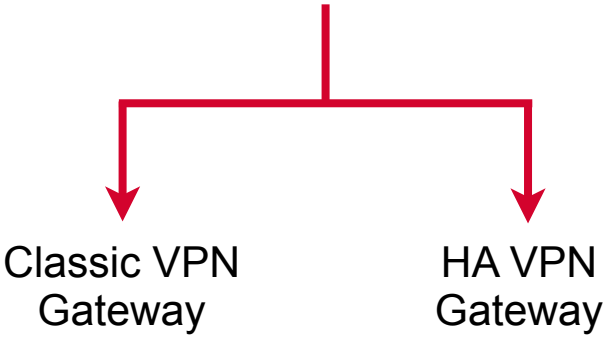
- Two VPN gateways
- One for cloud network, another for on-prem network

- Traffic encrypted at one gateway
- Decrypted at other gateway
- Keys need to be exchanged

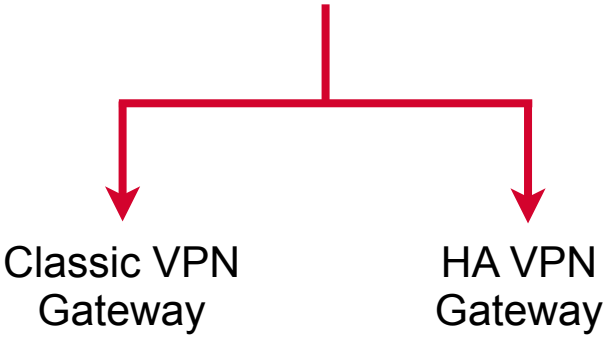
Cloud VPN


Cloud Network
On-prem Network
Encrypted VPN Tunnel
VPN
VPN
Cloud VPN Gateway
Peer VPN Gateway

VPN Gateways



VPN Gateways



Legacy offering - not recommended for new deployments.

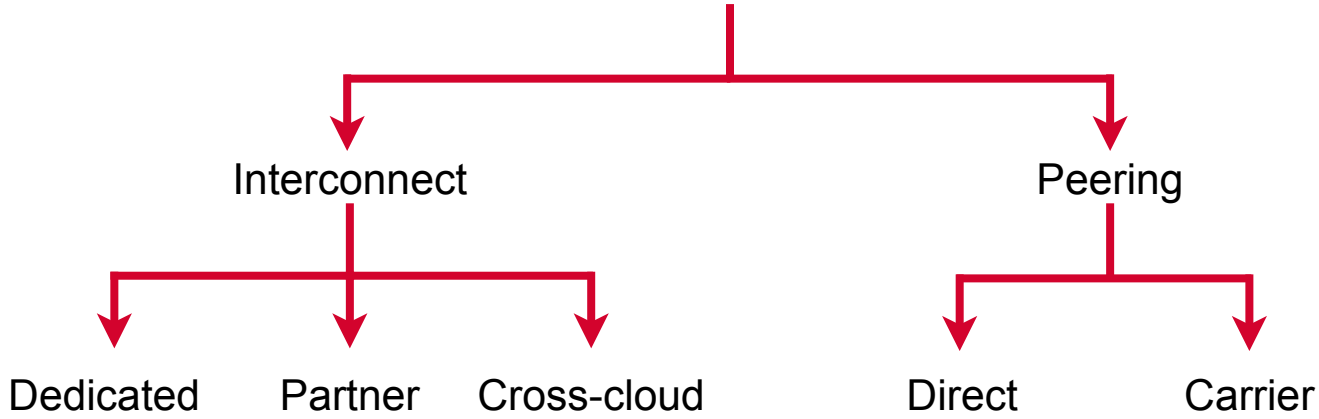
No redundancy - if a single gateway or tunnel fails connection is lost

Recommended option - has two separate interfaces with different IPs

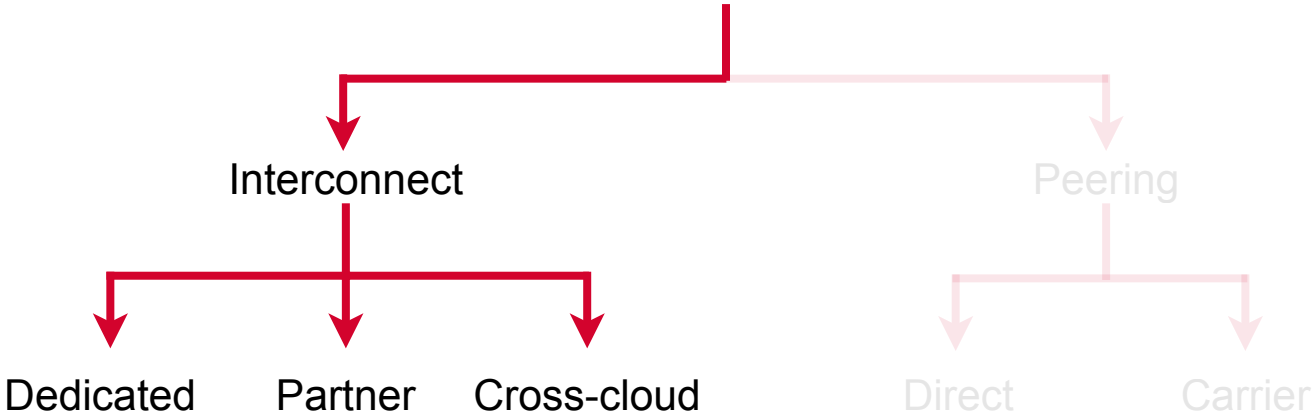
For redundancy configure two tunnels from on-prem to cloud. Traffic will automatically failover to the other if needed

Enterprise Connectivity

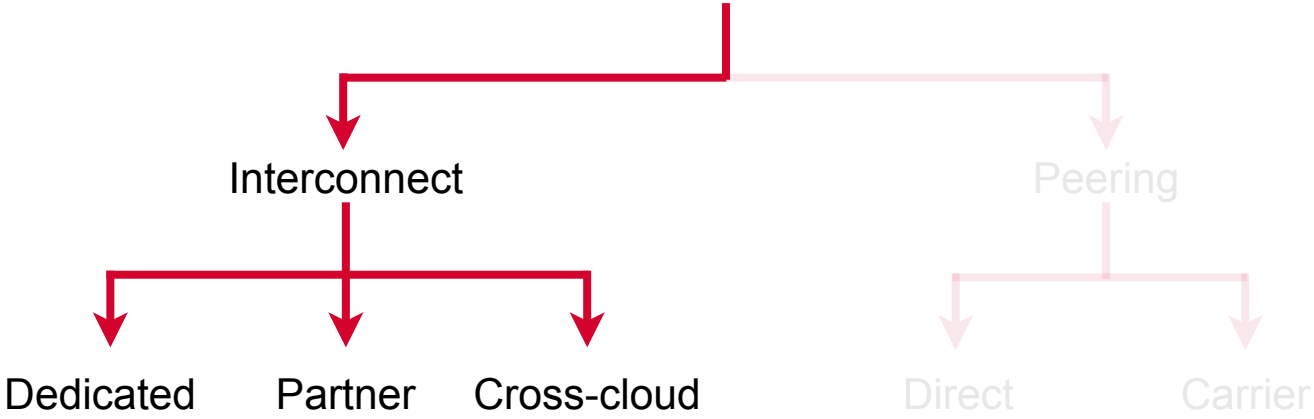
Enterprise Connectivity



Enterprise Connectivity



Enterprise Connectivity



Internal IP addresses in
RFC 1918 address space

With SLA

Traffic between your external network and Google network DOES NOT traverse the public internet

Resources on connected networks communicate via Internal IPs

Dedicated Interconnect

Configuration Property	Choice
Connection	Dedicated, direct connection to VPC networks
Access Type	Internal IP addresses in RFC 1918 address space
Capacity	10 Gbps or 100 Gbps connections
Other Considerations	**Must have connection in a Google supported colocation facility that supports the regions you want to connect to**

Partner Interconnect

Configuration Property	Choice
Connection	Dedicated Bandwidth, connection to VPC network through a service provider
Access Type	Internal IP addresses in RFC 1918 address space
Capacity	50Mbps - 50Gbps per connection
Other Considerations	**Service providers might have specific restrictions or requirements**

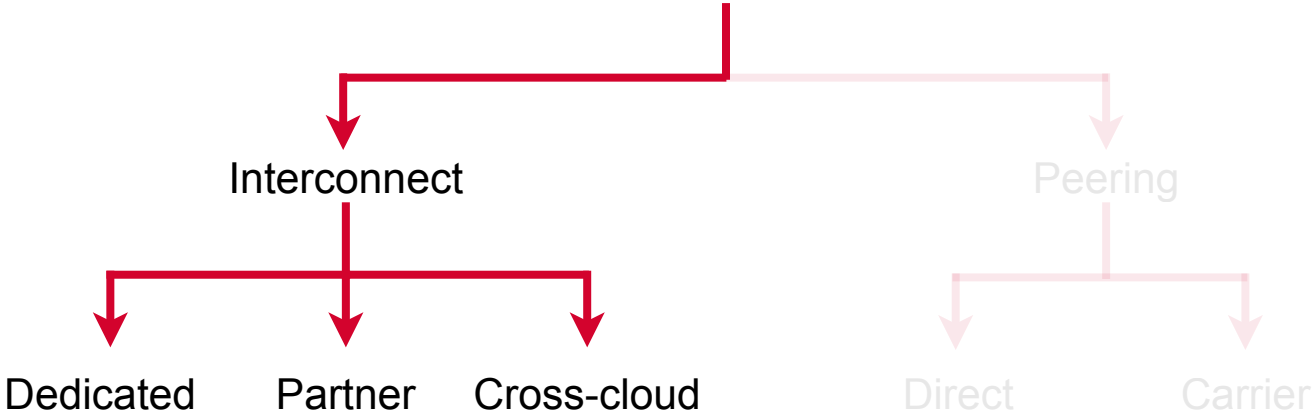
Cross-cloud Interconnect

- High-bandwidth dedicated connectivity between Google Cloud and another service provider
- Google will provision a dedicated physical connection
- Useful for:
  - Site-to-site data transfer
  - Multi-cloud strategy

Cross-cloud Interconnect

Configuration Property	Choice
Connection	Dedicated physical connection between Google Cloud and other cloud platform
Access Type	Internal IP addresses in RFC 1918 address space
Capacity	10 Gbps or 100Gbps
Other Considerations	**Supported cloud provides AWS, Azure, Oracle, Alibaba**

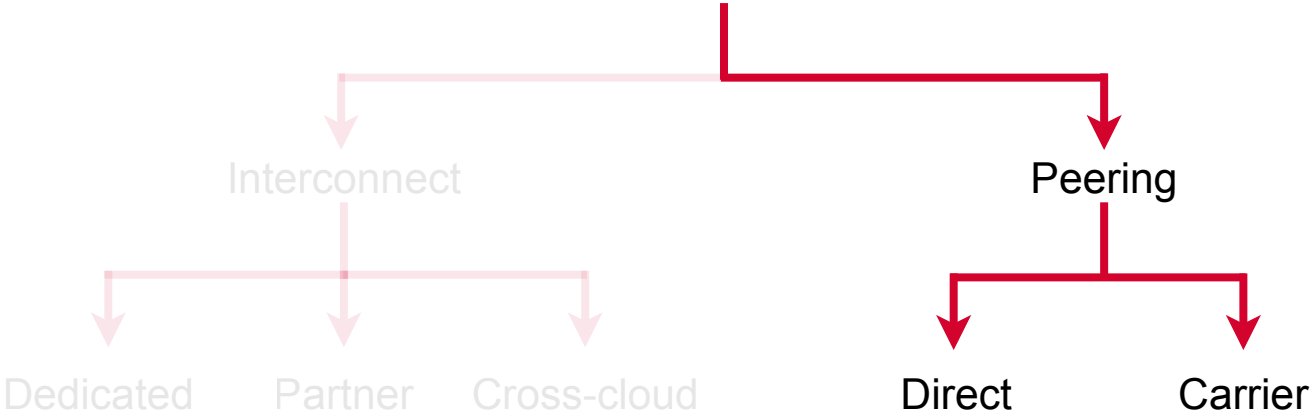
Enterprise Connectivity



Internal IP addresses in
RFC 1918 address space

With SLA

Enterprise Connectivity



Public IP addresses

No SLA

Direct Peering

Configuration Property	Choice
Connection	Provides direct access from your on-premises network to Google Workspace and Google APIs for the full suite of Google Cloud products.
Access Type	Public IP addresses
Other Considerations	**Connects to Google's edge network**

Carrier Peering

Configuration Property	Choice
Connection	Peering through service provider to access Google applications such as Google Workspace and to Google Cloud products that can be exposed through one or more public IP addresses.
Access Type	Public IP addresses
Other Considerations	**Connects to Google's edge network through a service provider. Requirements vary by partner**

Cloud Router

- Cloud Router is a fully distributed and managed Google Cloud service that **dynamically manages routing tables**
- Uses the Border **Gateway Protocol (BGP)** to exchange routes between Google Cloud and on-premise networks
- Allows for **automatic updation** when network changes occur
- Used with Cloud Interconnect and Cloud VPN

Cloud NAT

- Managed Google service that allows resources without public IP addresses to initiate outbound connections to the internet
  - VMs with no external IP addresses
  - GKE nodes on a private cluster
- Manage and monitor outbound internet access from a central point
- Use a predictable set of public IPs that can be allowlisted by external services.

Networking

A financial services company is extending its on-premises data center into Google Cloud to create a hybrid environment. They plan to run application servers on Google Cloud that need to communicate directly with a database server remaining on-premises, using only private, internal IP addresses (RFC 1918).

The company requires a networking solution that provides continuous connectivity even when an existing link is down.

Which combination of services should they implement to achieve this?

A. Use a **Global External HTTP/S Load Balancer** with an on-premises server as an Internet NEG backend, and a second load balancer for path diversity.

B. Establish a **Dedicated Interconnect** for a private physical path. For path diversity, configure **Direct Peering** to Google's network edge.

C. Set up a **Shared VPC** to link the on-premises data center and the Google Cloud VPC, and use **Cloud Router** to manage routes.

D. Establish a **Dedicated Interconnect** for a high-bandwidth, private physical path. Concurrently, configure a **Cloud VPN** tunnel to provide an encrypted path over the public internet.

Networking

A financial services company is extending its on-premises data center into Google Cloud to create a hybrid environment. They plan to run application servers on Google Cloud that need to communicate directly with a database server remaining on-premises, using only private, internal IP addresses (RFC 1918).

The company requires a networking solution that provides continuous connectivity even when an existing link is down.

Which combination of services should they implement to achieve this?

A. Use a **Global External HTTP/S Load Balancer** with an on-premises server as an Internet NEG backend, and a second load balancer for path diversity.

B. Establish a **Dedicated Interconnect** for a private physical path. For path diversity, configure **Direct Peering** to Google's network edge.

C. Set up a **Shared VPC** to link the on-premises data center and the Google Cloud VPC, and use **Cloud Router** to manage routes.

D. Establish a Dedicated Interconnect for a high-bandwidth, private physical path. Concurrently, configure a Cloud VPN tunnel to provide an encrypted path over the public internet.

O'REILLY®

Cloud CDN

Cloud CDN

Cloud CDN (Content Delivery Network) uses Google's global edge network (Points of Presence or PoPs) to serve content closer to users, which accelerates your websites and applications.

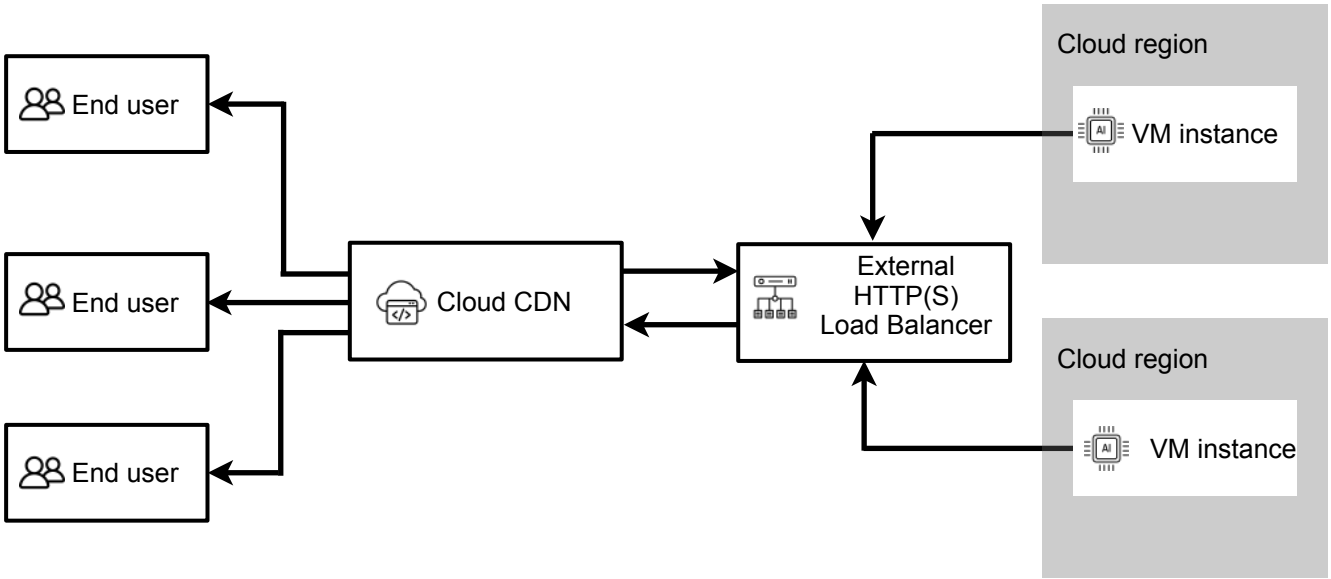
*Google Edge Network consists of numerous edge locations that are spread across various cities and countries globally. These edge locations are situated closer to users than Google’s central data centers, reducing latency by ensuring that users' data and requests travel shorter distances.

Cloud CDN

Usually sits in front of a load balancer and caches content from various types of backends.

Backends referred to as origin servers.

Cloud CDN



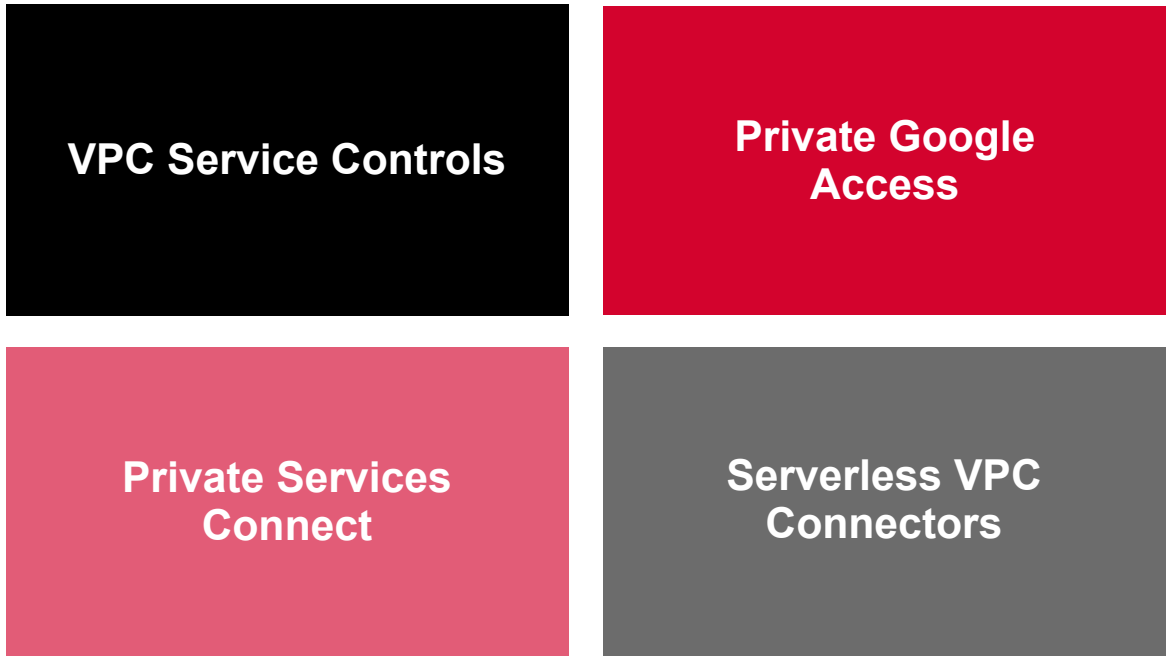
The Cloud CDN cache stores and manages content so that future requests for that content can be served faster. The cached content is a copy of cacheable content that is stored on origin servers.

Improve Performance of Cache

- Cache keys used to uniquely identify cached assets
- Ignore unnecessary query parameters in URLs while storing cached content
  - e.g. ?utm_source=google
- Normalize cache keys
  - HTTP vs. HTTPS
  - Hostnames (if they serve the same content)


O'REILLY®
VPC Secure Access Features

VPC Secure Access Features



VPC Secure Access Features

VPC Service Controls

Private Google Access

Private Services Connect

Serverless VPC Access

VPC Secure Access Features

VPC Service Controls

VPC Service Controls

Help protect against accidental or targeted data exfiltration risks from Google Cloud services such as Cloud Storage and BigQuery.

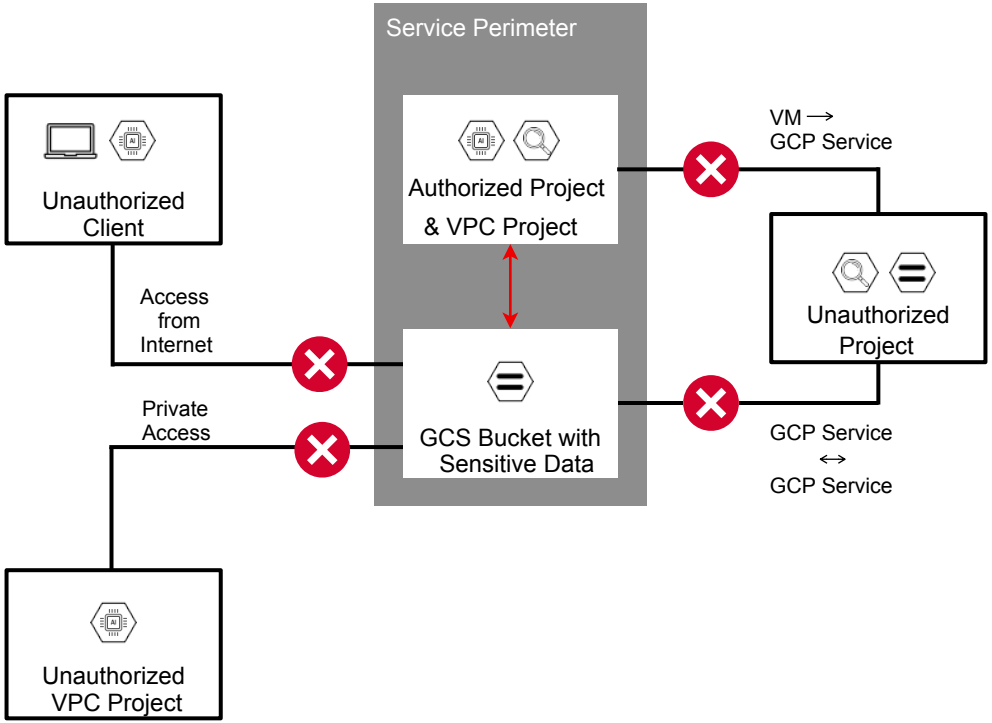
Creates service perimeters that protect the resources or data that you specify

Service Perimeter

A **service perimeter** creates a security boundary around Google Cloud resources.

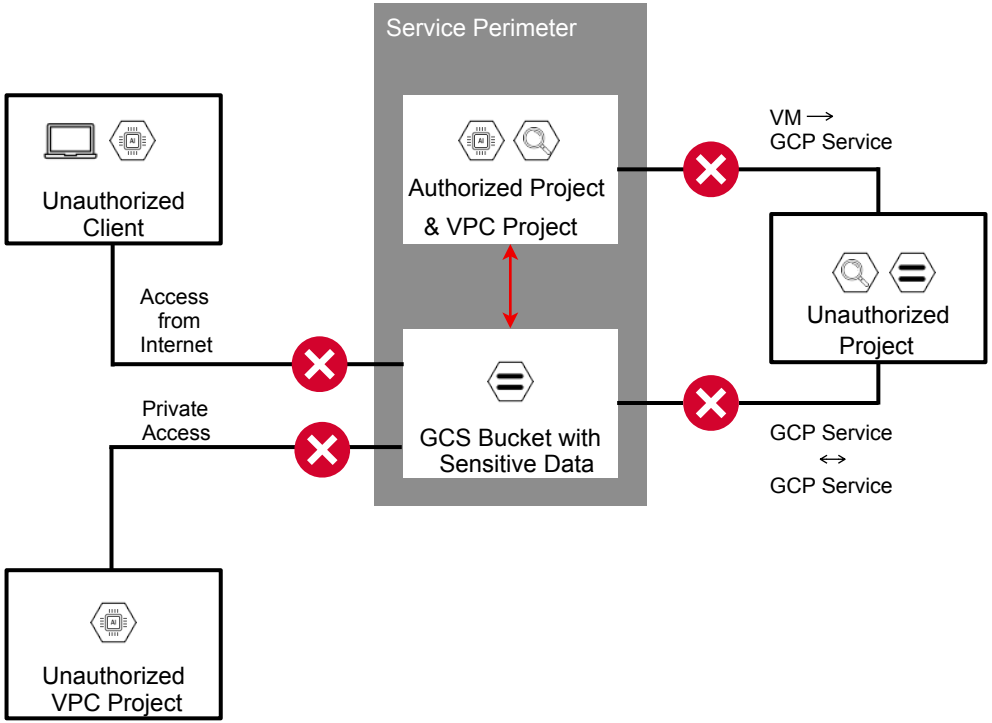
A service perimeter allows free communication within the perimeter but, by default, blocks communication to Google Cloud services across the perimeter.

Define Security Perimeters



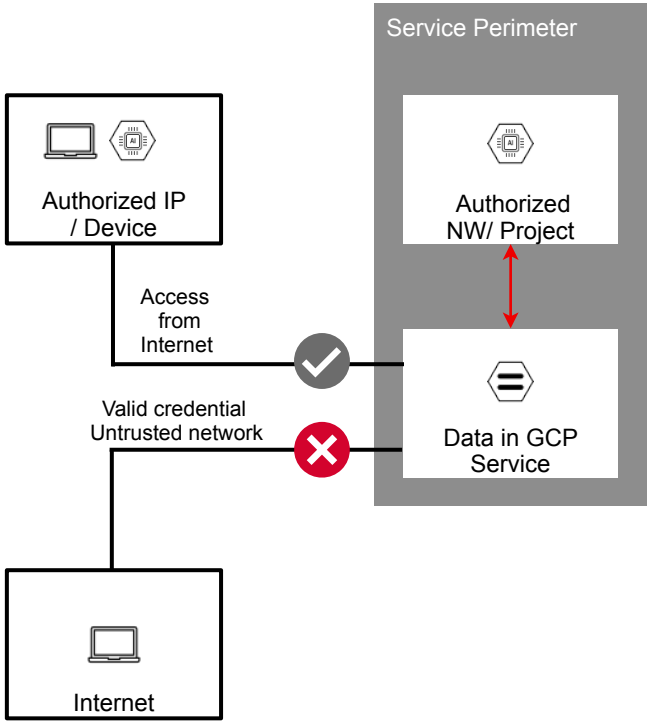
Include resources with sensitive data
Include services with access to those resources

Control Data Movement In and Out of Perimeter



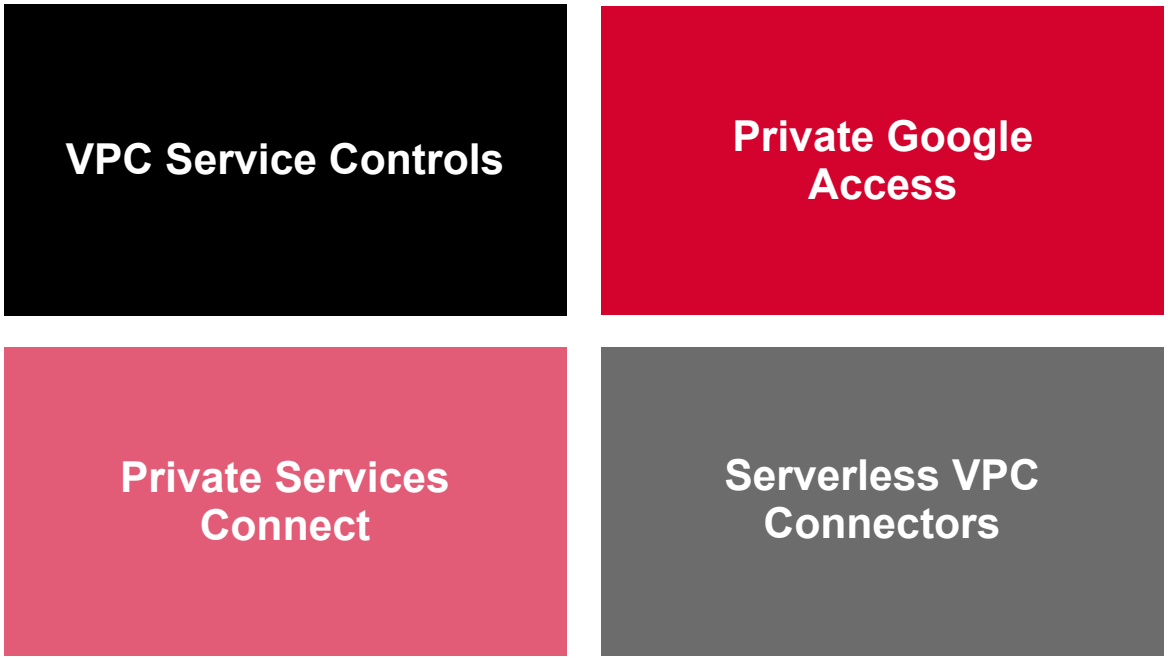
Data cannot be copied to unauthorized resources outside the perimeter
Data exchange across the perimeter controlled by ingress and egress rules

Context Aware Access



Based on identity of the user, device state, network origin, other context signals

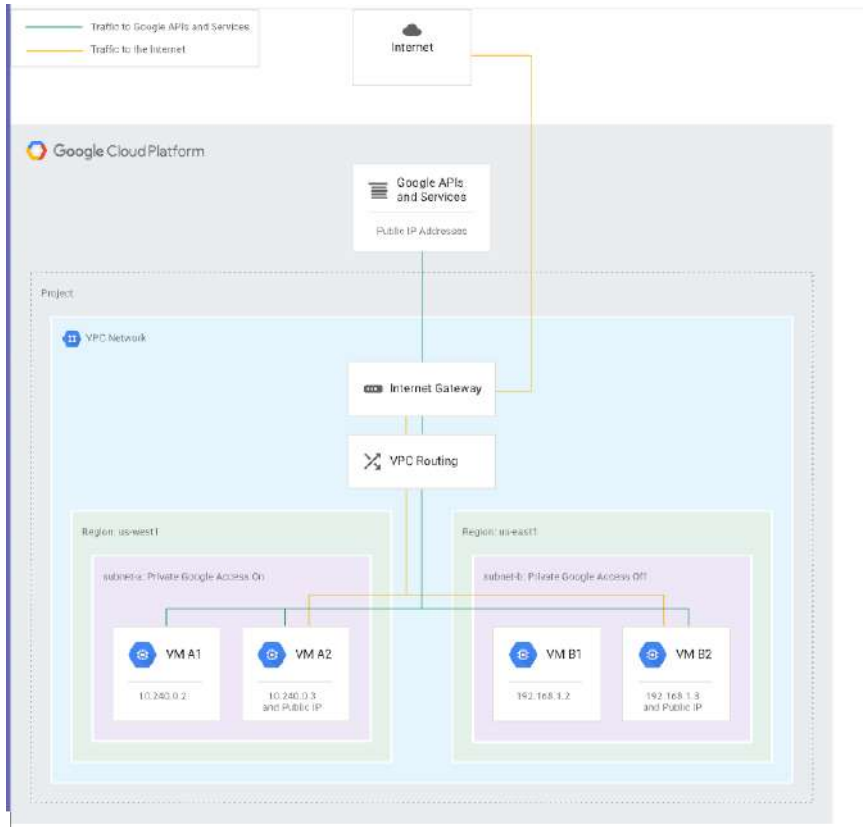
VPC Secure Access Features



Private Google Access

Allows resources **without** public IP addresses (like private GKE nodes or Compute Engine VMs) to reach the *public endpoints* of Google APIs and services (like Cloud Storage, BigQuery, or Artifact Registry).

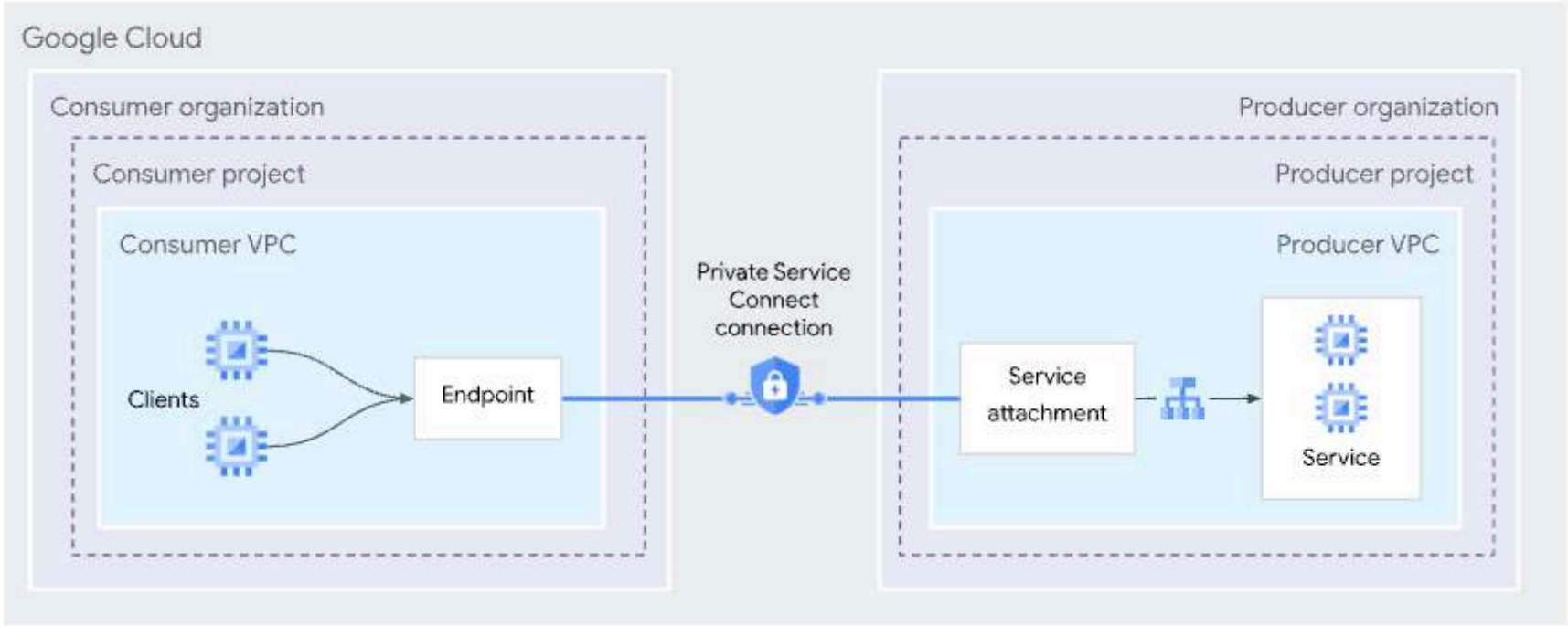
Private Google Access



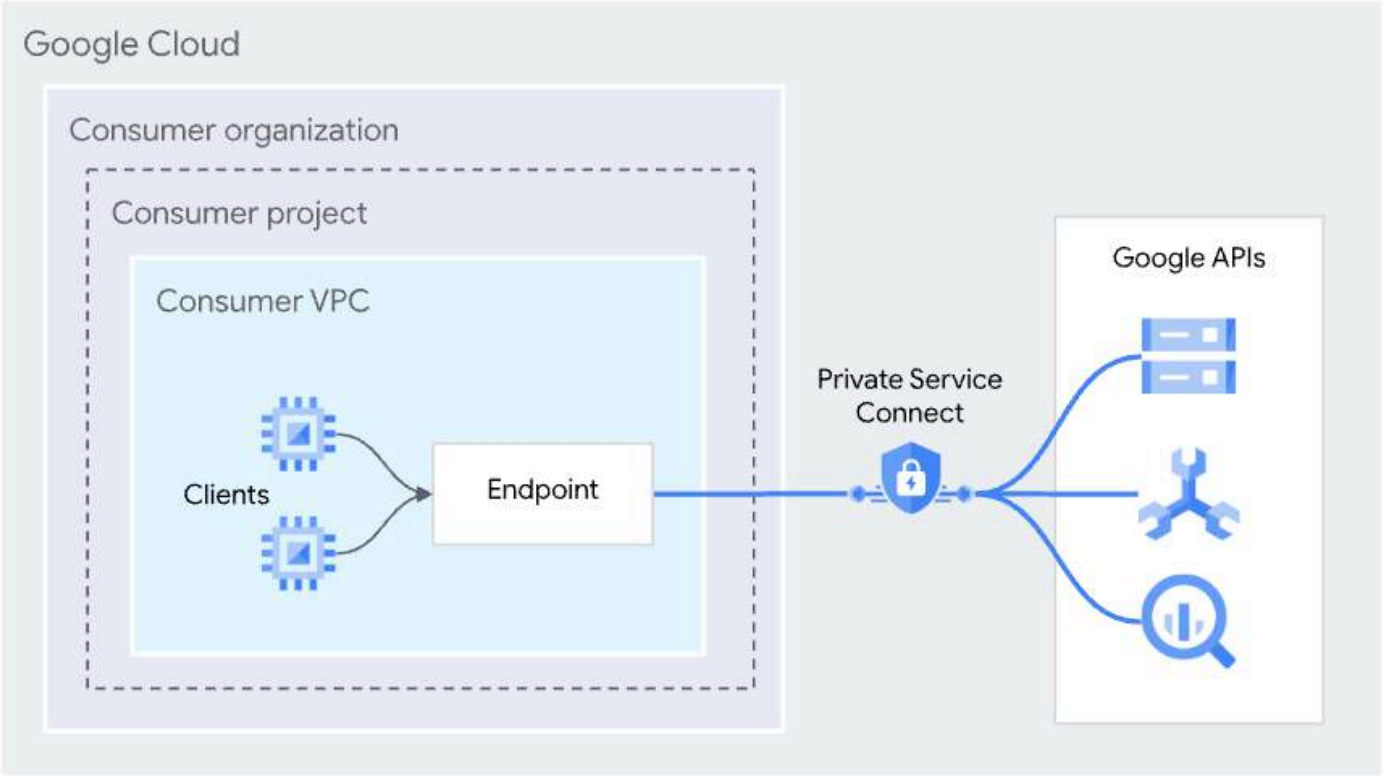
Private Service Connect

Allows you to access a service (whether it's a Google service like Cloud SQL, a third-party SaaS product, or even a service you built in another VPC) using a **private IP address that you own** within your VPC. It makes an external service feel like it's a native part of your own network.

Private Service Connect



Private Service Connect



How do server less services connect to a VPC?

Using VPC Connectors

Serverless Environments

- Serverless environments like those for Cloud Run, App Engine, Cloud Functions are **outside your VPC**
- **How can your Cloud Function connect to your Cloud SQL database using internal IPs?**
- The Serverless VPC Connector is a bridge that connects the server less environments to your VPC
  - Allows access to resources on your VPC using internal IP addresses
  - More secure, lower latency

Networking

A stock photography company, "PixelPerfect," hosts its library of high-resolution images in a Cloud Storage bucket. Their website, running on a fixed set of Compute Engine instances, serves these images to a global user base of graphic designers.

Customers are reporting that downloading the most popular and featured images of the week is frequently slow and sometimes times out, especially during peak business hours in their respective regions. You need to accelerate the delivery of these popular images and improve the reliability of the download experience for all users.

What should you do?

A. Write a script that syncs the Cloud Storage bucket to a Persistent Disk attached to each Compute Engine VM. Modify the application to serve the images directly from the VMs' local disks.

B. Provision a **Memorystore for Redis** instance. For popular images, load the raw image data into Redis as byte strings. Update the application to check Redis first before fetching from Cloud Storage.

C. Create a **Bigtable** table to store the image data. Rewrite the application to stream the images from Bigtable rows to reduce the load on Cloud Storage for popular files.

D. Set up a **Global External HTTP/S Load Balancer**. Configure one of its backend services to be the **Cloud Storage bucket** containing the images. Ensure that **Cloud CDN** is enabled for this backend service.

Networking

A stock photography company, "PixelPerfect," hosts its library of high-resolution images in a Cloud Storage bucket. Their website, running on a fixed set of Compute Engine instances, serves these images to a global user base of graphic designers.

Customers are reporting that downloading the most popular and featured images of the week is frequently slow and sometimes times out, especially during peak business hours in their respective regions. You need to accelerate the delivery of these popular images and improve the reliability of the download experience for all users.

What should you do?

A. Write a script that syncs the Cloud Storage bucket to a Persistent Disk attached to each Compute Engine VM. Modify the application to serve the images directly from the VMs' local disks.

B. Provision a **Memorystore for Redis** instance. For popular images, load the raw image data into Redis as byte strings. Update the application to check Redis first before fetching from Cloud Storage.

C. Create a **Bigtable** table to store the image data. Rewrite the application to stream the images from Bigtable rows to reduce the load on Cloud Storage for popular files.

D. Set up a Global External HTTP/S Load Balancer. Configure one of its backend services to be the Cloud Storage bucket containing the images. Ensure that Cloud CDN is enabled for this backend service.

Networking

A retail company is building a hybrid application where a new Cloud Function on GCP is used for real-time inventory checks. This function needs to query a high-performance Redis cache that resides in their on-premises data center.

The company's Google Cloud VPC is already connected to their on-premises network via a stable Cloud VPN tunnel. The Cloud Function must be able to send requests to the private IP address of the on-premises Redis cache.

What must be configured in Google Cloud to enable this communication path?

A. Modify the Cloud VPN tunnel's firewall rules to explicitly allow traffic originating from the public IP ranges of Google Cloud Functions.

B. Assign a public static IP address to the on-premises Redis cache using a NAT gateway and have the Cloud Function connect to it over the public internet.

C. Configure Private Google Access for the on-premises network, allowing on-premises hosts to communicate privately with Google services like Cloud Functions.

D. Create a Serverless VPC Access connector. Configure the Cloud Function to route its traffic through this connector into the VPC, allowing the traffic to then traverse the Cloud VPN tunnel.

Networking

A retail company is building a hybrid application where a new Cloud Function on GCP is used for real-time inventory checks. This function needs to query a high-performance Redis cache that resides in their on-premises data center.

The company's Google Cloud VPC is already connected to their on-premises network via a stable Cloud VPN tunnel. The Cloud Function must be able to send requests to the private IP address of the on-premises Redis cache.

What must be configured in Google Cloud to enable this communication path?

A. Modify the Cloud VPN tunnel's firewall rules to explicitly allow traffic originating from the public IP ranges of Google Cloud Functions.

B. Assign a public static IP address to the on-premises Redis cache using a NAT gateway and have the Cloud Function connect to it over the public internet.

C. Configure Private Google Access for the on-premises network, allowing on-premises hosts to communicate privately with Google services like Cloud Functions.

D. Create a Serverless VPC Access connector. Configure the Cloud Function to route its traffic through this connector into the VPC, allowing the traffic to then traverse the Cloud VPN tunnel.

Networking

A healthcare research institute stores sensitive patient data in Google Cloud Storage buckets within a dedicated GCP project. Researchers access this data for analysis from Compute Engine VMs inside the same project's Virtual Private Cloud (VPC).

The security team must enforce a strict security perimeter to prevent data exfiltration. They need to block any attempt by a compromised VM or a malicious user to copy data from the project's secure buckets to a public bucket or any other unauthorized Google Cloud location, while still allowing the VMs to access the required Cloud Storage APIs for their work.

What is the most effective Google Cloud solution to meet these requirements?

A. Create a custom IAM role that only grants storage.objects.get permission. Assign this role to researchers and remove all broader storage roles to enforce least privilege.

B. Enforce the use of Customer-Managed Encryption Keys (CMEK) for all Cloud Storage buckets. Only grant key decryption permissions to authorized service accounts.

C. Define a VPC Service Controls perimeter that includes the project and the required Google APIs (e.g., Cloud Storage API). Ensure Private Google Access is enabled on the VPC's subnets.

D. Configure the project's VPC firewall rules to deny all egress traffic to the internet. Enable Private Google Access on the subnet to allow VMs to continue reaching Google APIs.

Networking

A healthcare research institute stores sensitive patient data in Google Cloud Storage buckets within a dedicated GCP project. Researchers access this data for analysis from Compute Engine VMs inside the same project's Virtual Private Cloud (VPC).

The security team must enforce a strict security perimeter to prevent data exfiltration. They need to block any attempt by a compromised VM or a malicious user to copy data from the project's secure buckets to a public bucket or any other unauthorized Google Cloud location, while still allowing the VMs to access the required Cloud Storage APIs for their work.

What is the most effective Google Cloud solution to meet these requirements?

A. Create a custom IAM role that only grants storage.objects.get permission. Assign this role to researchers and remove all broader storage roles to enforce least privilege.

B. Enforce the use of Customer-Managed Encryption Keys (CMEK) for all Cloud Storage buckets. Only grant key decryption permissions to authorized service accounts.

C. Define a VPC Service Controls perimeter that includes the project and the required Google APIs (e.g., Cloud Storage API). Ensure Private Google Access is enabled on the VPC's subnets.

D. Configure the project's VPC firewall rules to deny all egress traffic to the internet. Enable Private Google Access on the subnet to allow VMs to continue reaching Google APIs.

Networking

A financial company is deploying applications on a private Google Kubernetes Engine (GKE) cluster to maximize security. The GKE nodes in this cluster have been provisioned with no external IP addresses and cannot be reached directly from the internet.

However, to launch new applications, these nodes must be able to pull container images from public registries like Docker Hub, which are located on the internet. You need to enable this outbound connectivity for the GKE nodes without assigning them public IPs.

What is the recommended and most scalable Google Cloud solution?

A. Enable Private Google Access on the GKE cluster's subnet to allow the nodes to reach public endpoints on the internet.

B. Deploy a Compute Engine VM with a public IP to act as a forward proxy server. Configure the GKE nodes to route all outbound traffic through this proxy.

C. Provision a Cloud NAT gateway for the VPC network and configure it to apply to the subnet where the private GKE nodes reside.

D. Configure an Internal TCP/UDP Load Balancer within the VPC and create a firewall rule that allows it to forward traffic to the internet.

Networking

A financial company is deploying applications on a private Google Kubernetes Engine (GKE) cluster to maximize security. The GKE nodes in this cluster have been provisioned with no external IP addresses and cannot be reached directly from the internet.

However, to launch new applications, these nodes must be able to pull container images from public registries like Docker Hub, which are located on the internet. You need to enable this outbound connectivity for the GKE nodes without assigning them public IPs.

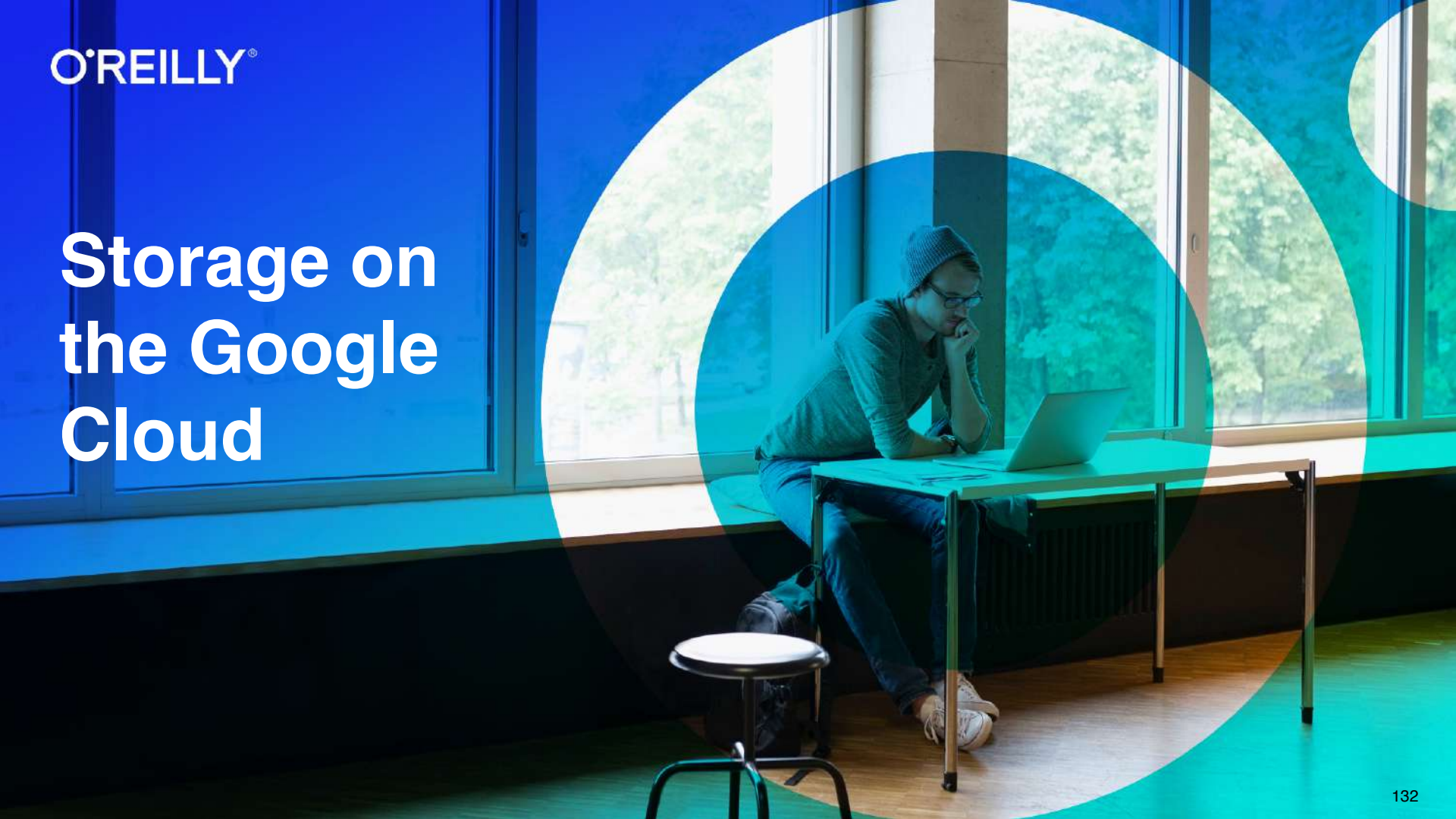
What is the recommended and most scalable Google Cloud solution?

A. Enable Private Google Access on the GKE cluster's subnet to allow the nodes to reach public endpoints on the internet.

B. Deploy a Compute Engine VM with a public IP to act as a forward proxy server. Configure the GKE nodes to route all outbound traffic through this proxy.

C. Provision a Cloud NAT gateway for the VPC network and configure it to apply to the subnet where the private GKE nodes reside.

D. Configure an Internal TCP/UDP Load Balancer within the VPC and create a firewall rule that allows it to forward traffic to the internet.


O'REILLY®
Storage on the Google Cloud

Choices in Computing

Compute

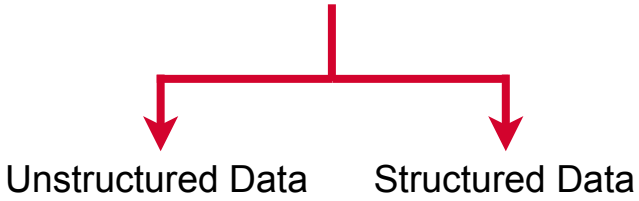
Where is code executed and how?

Storage

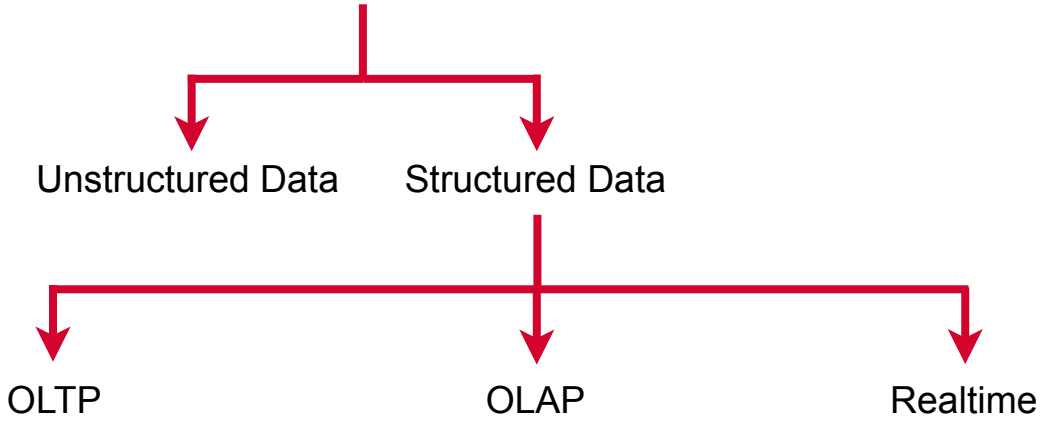
Where is data stored?

Networking, logging, are choices made after this fundamental decision

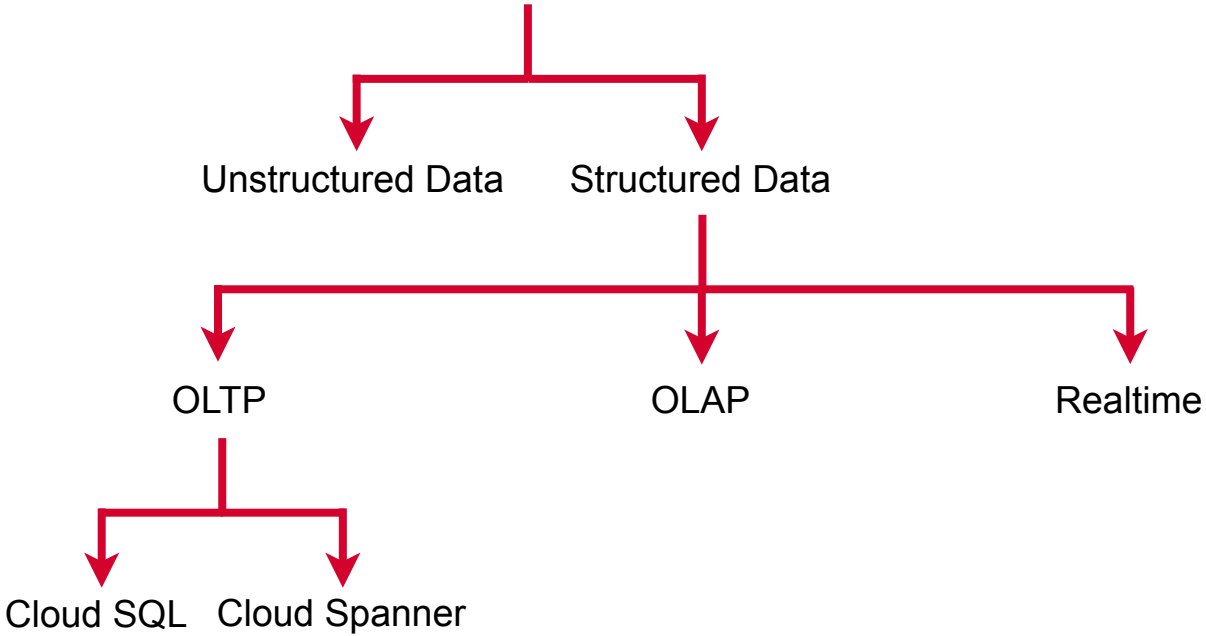
Storage Technologies



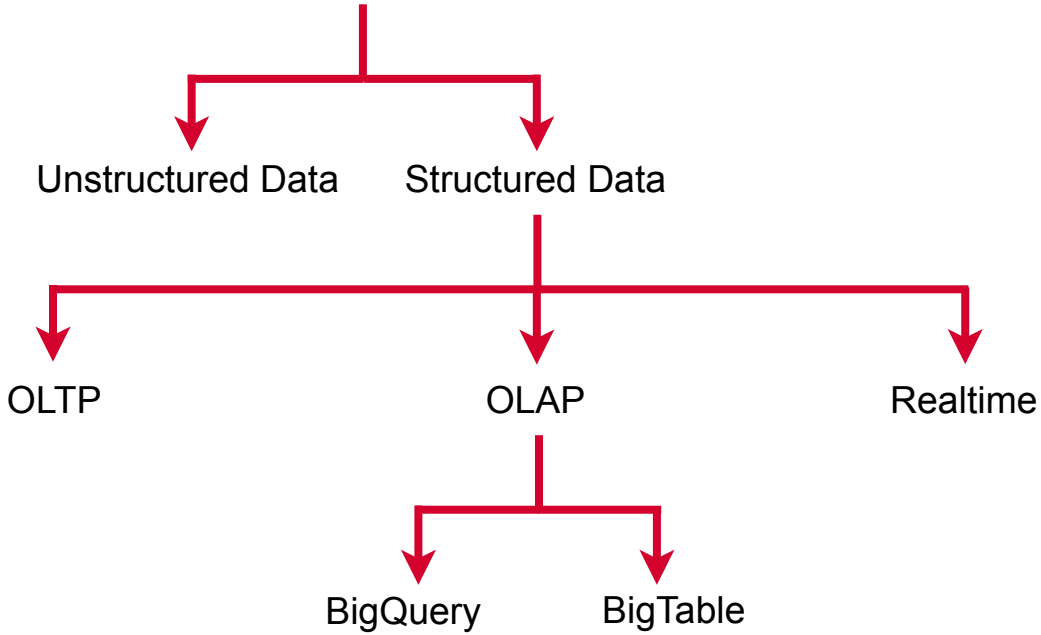
Storage Technologies



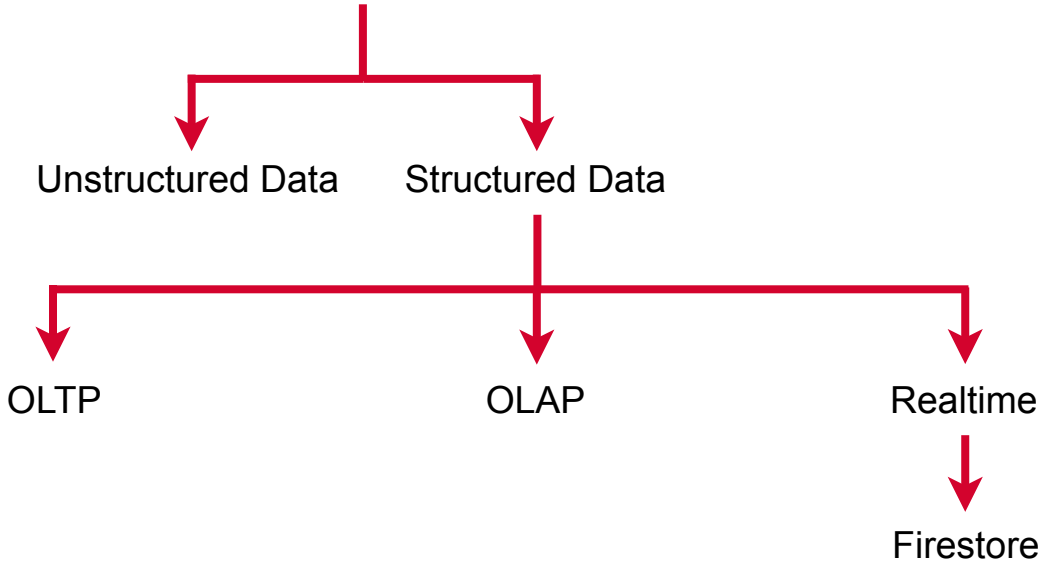
Storage Technologies



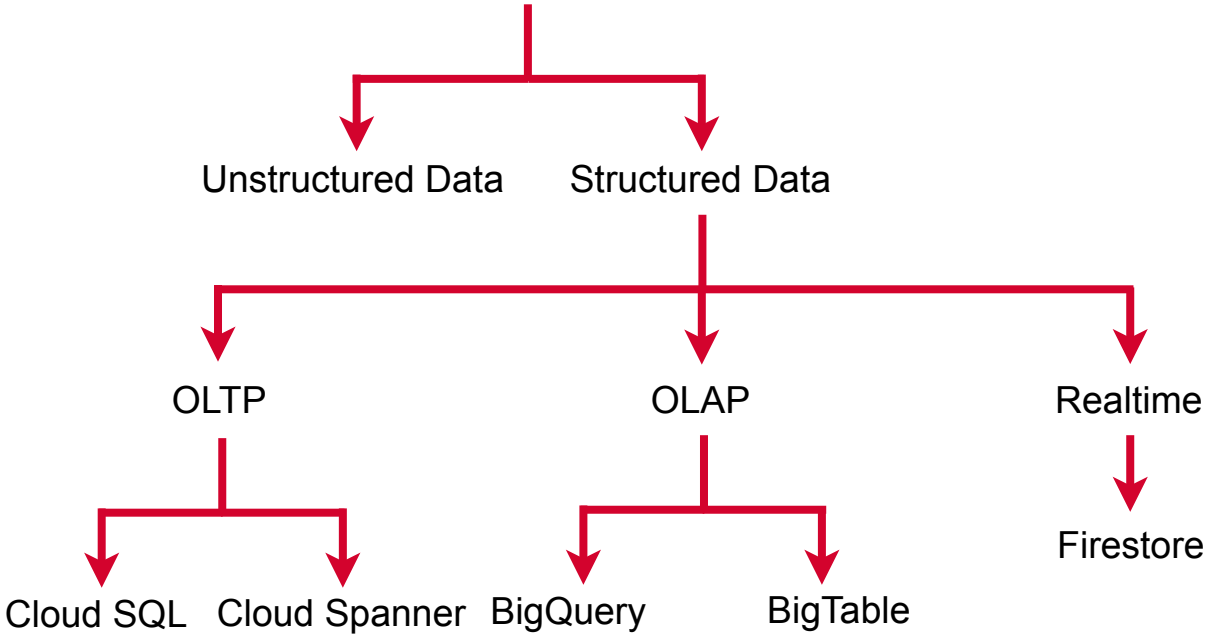
Storage Technologies



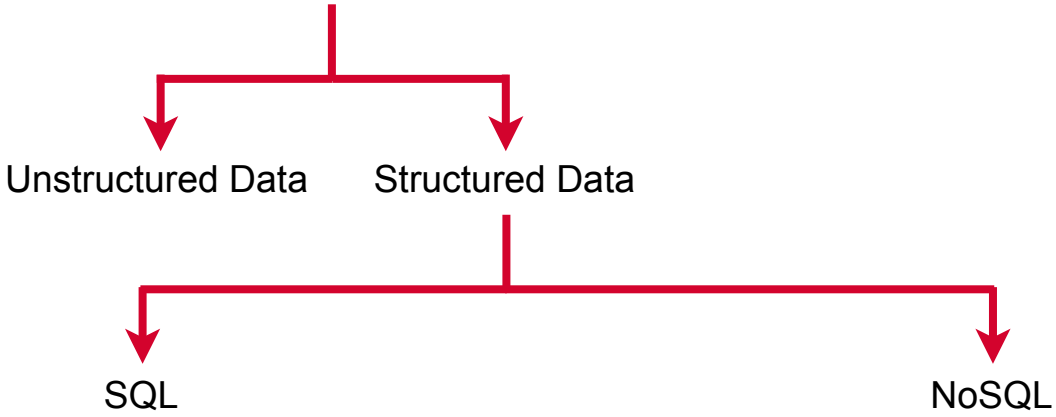
Storage Technologies



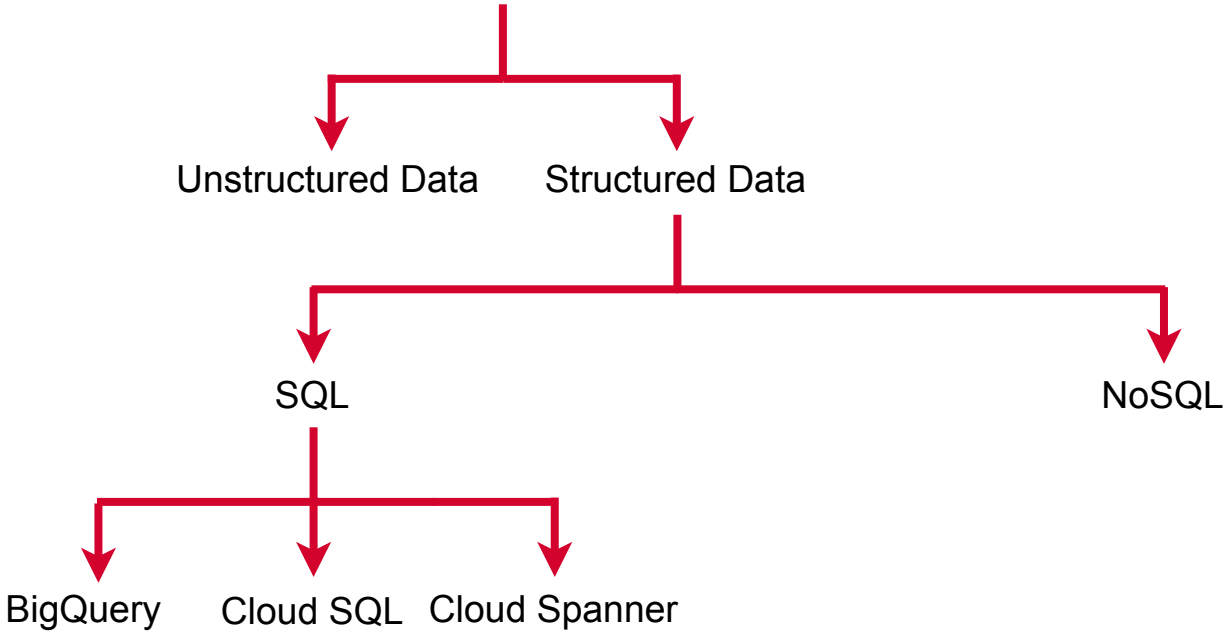
Storage Technologies



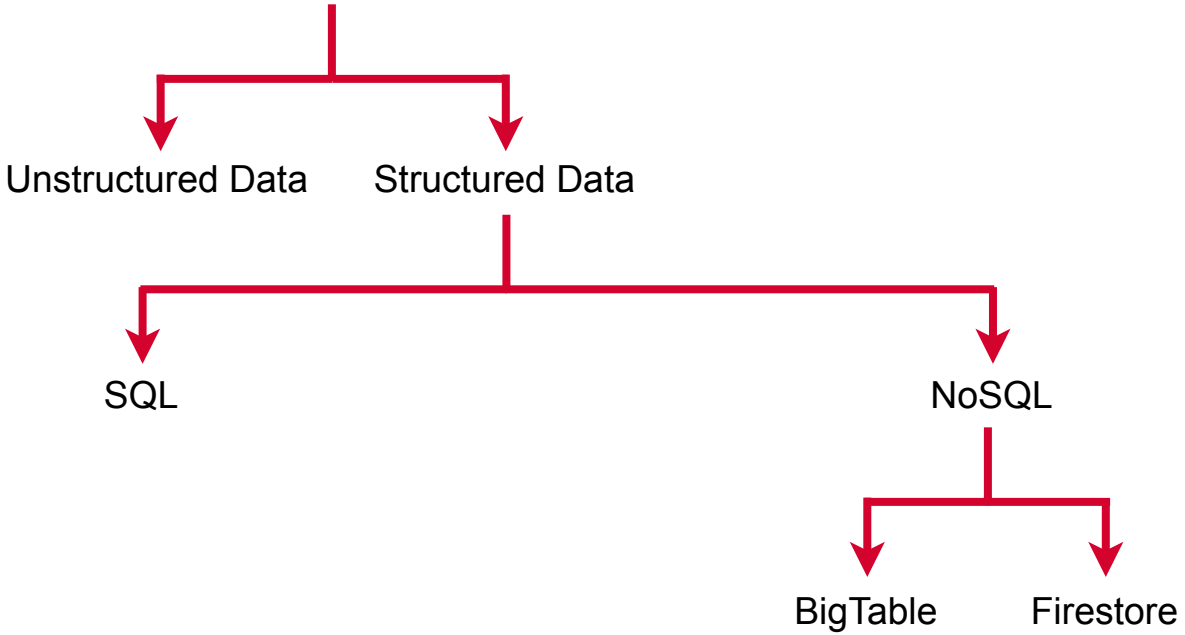
Storage Technologies



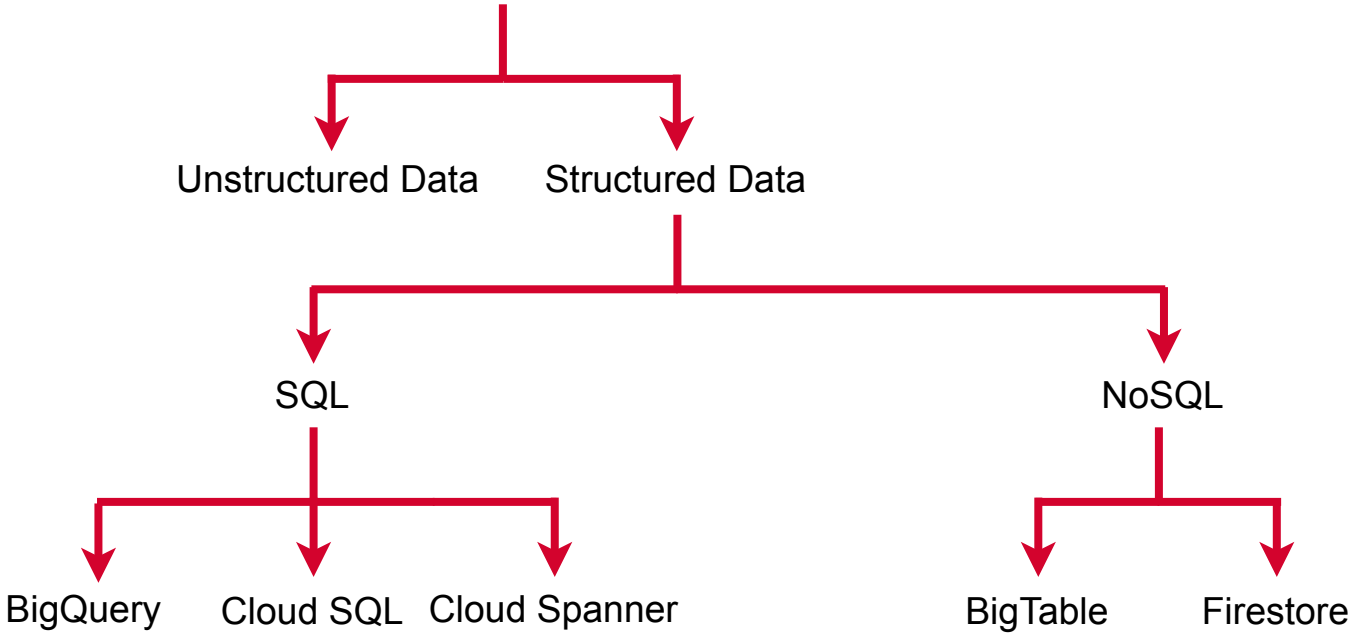
Storage Technologies



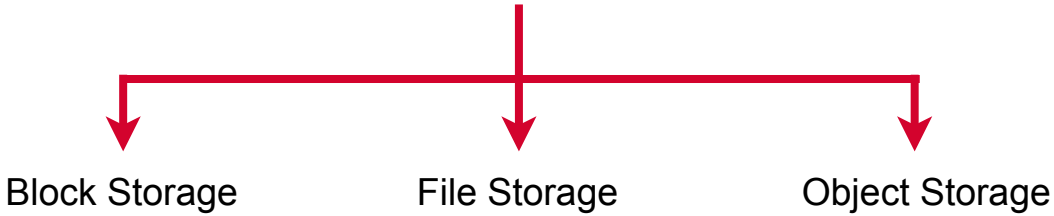
Storage Technologies



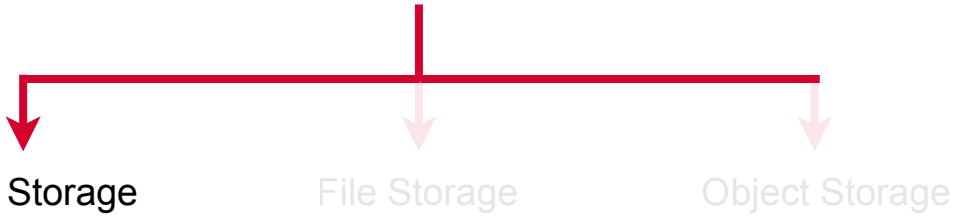
Storage Technologies



Unstructured Data



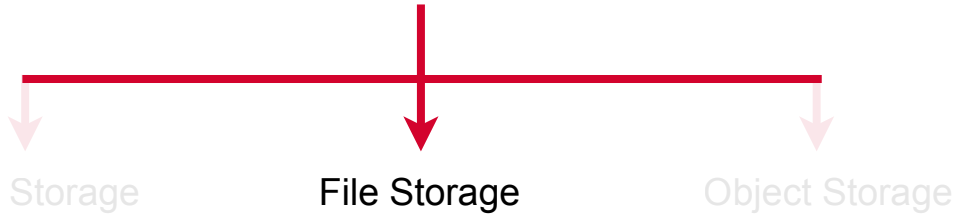
Persistent Disks



Physically addressable storage accessed from compute - data split into uniform blocks

High performance read and write access at the block level

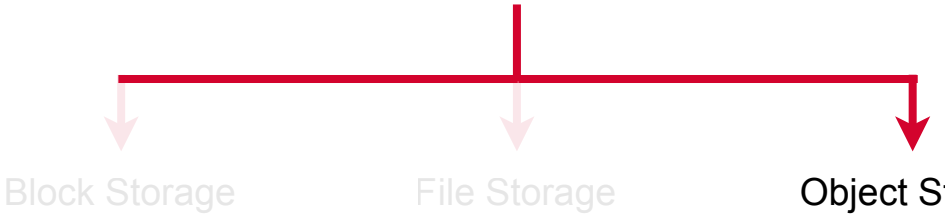
Cloud Filestore



Stores data as a hierarchy of files within directories

Shared concurrent access from multiple machines

Google Cloud Storage



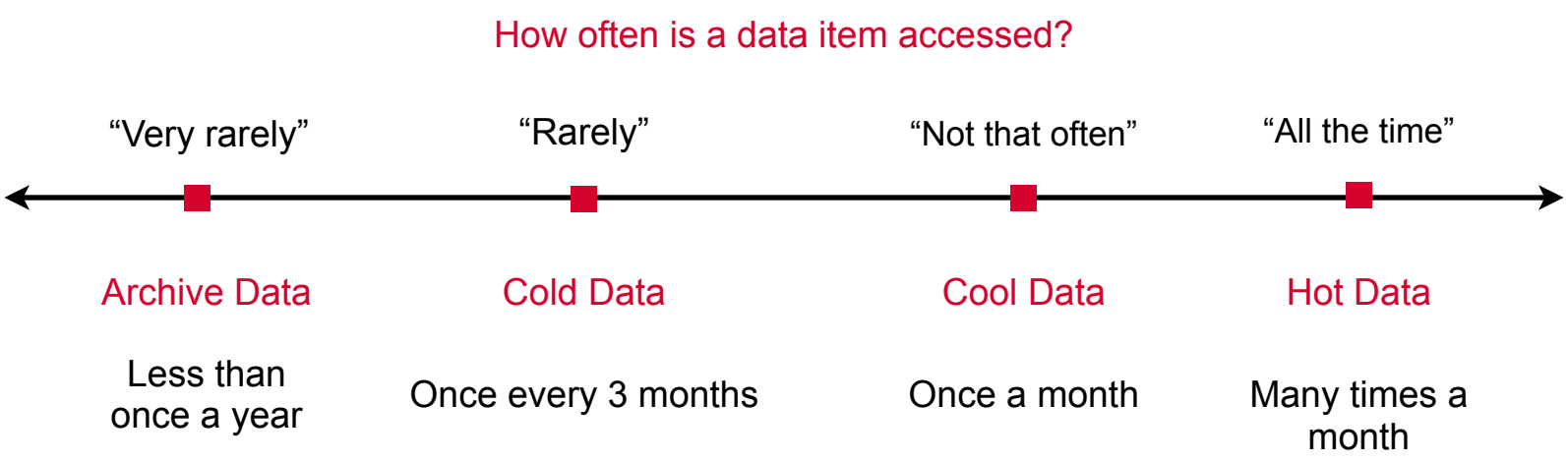
Logically addressable storage accessed from compute or by human users


O'REILLY®
Google
Cloud
Storage

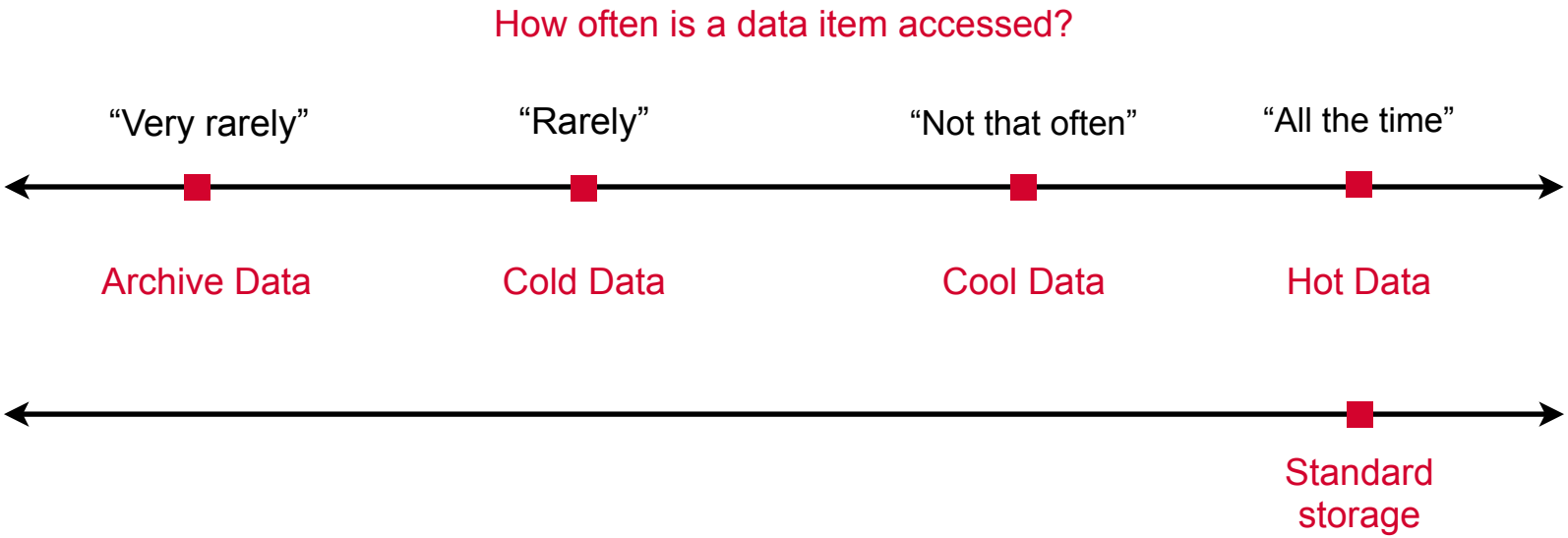
GCS Storage Classes

How often is a data item accessed?

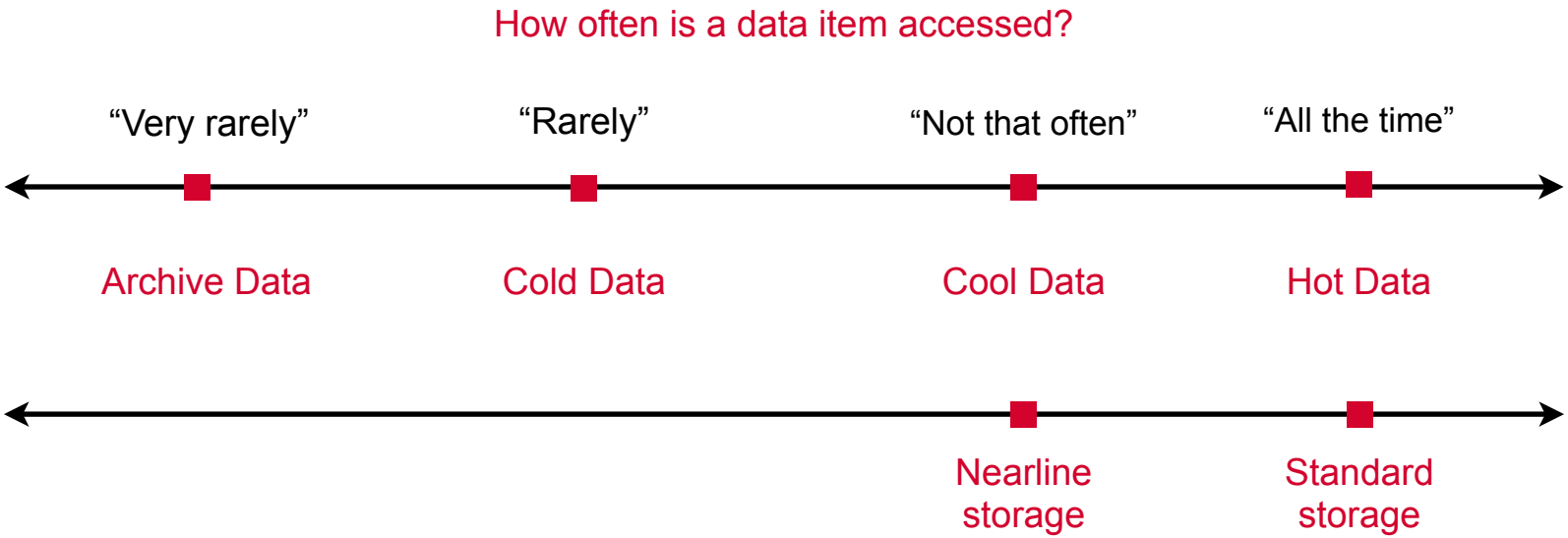
GCS Storage Classes



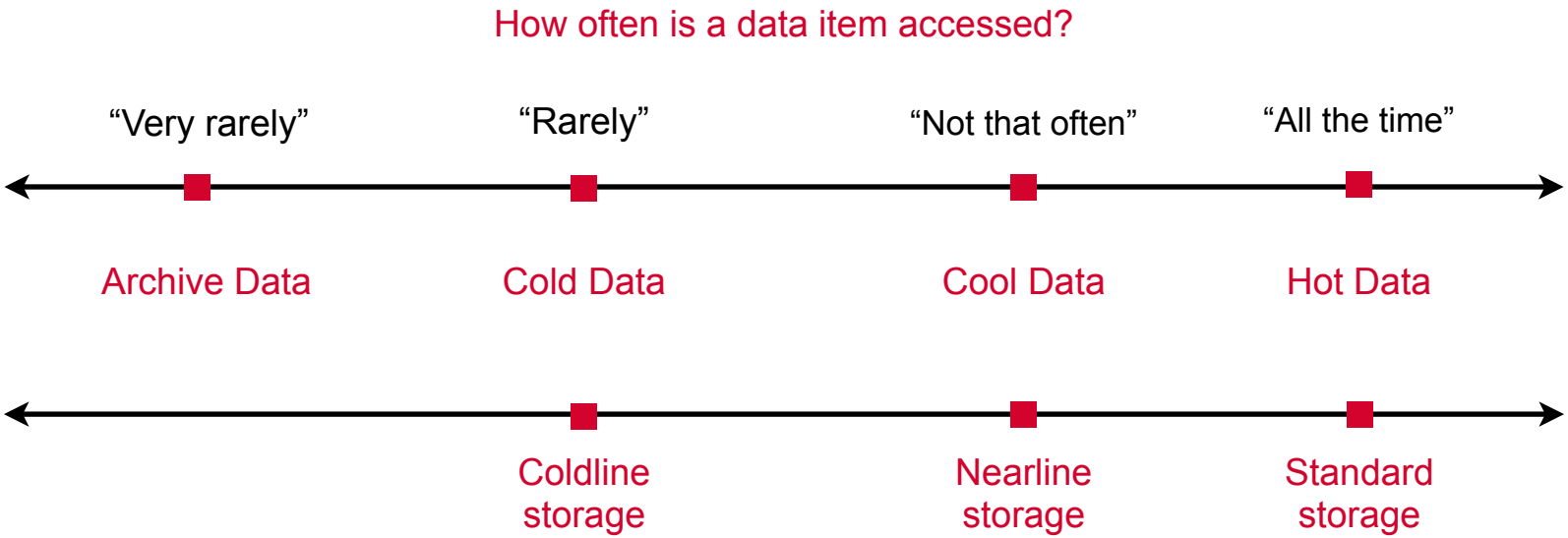
GCS Storage Classes



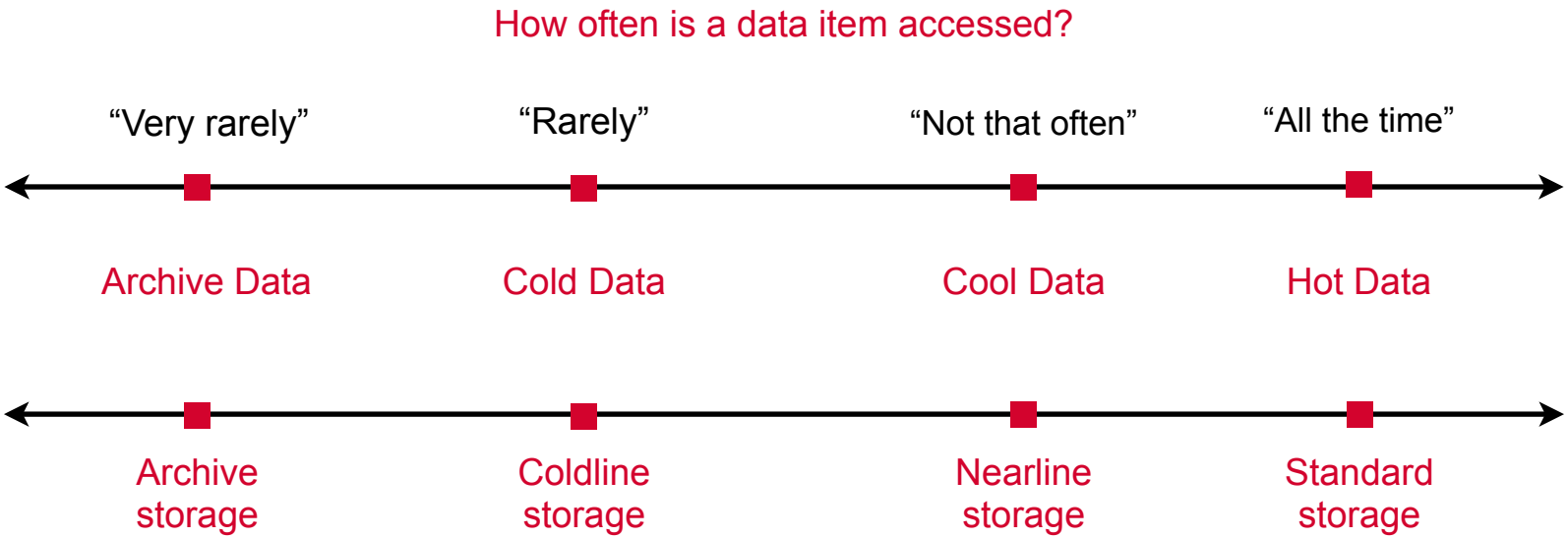
GCS Storage Classes



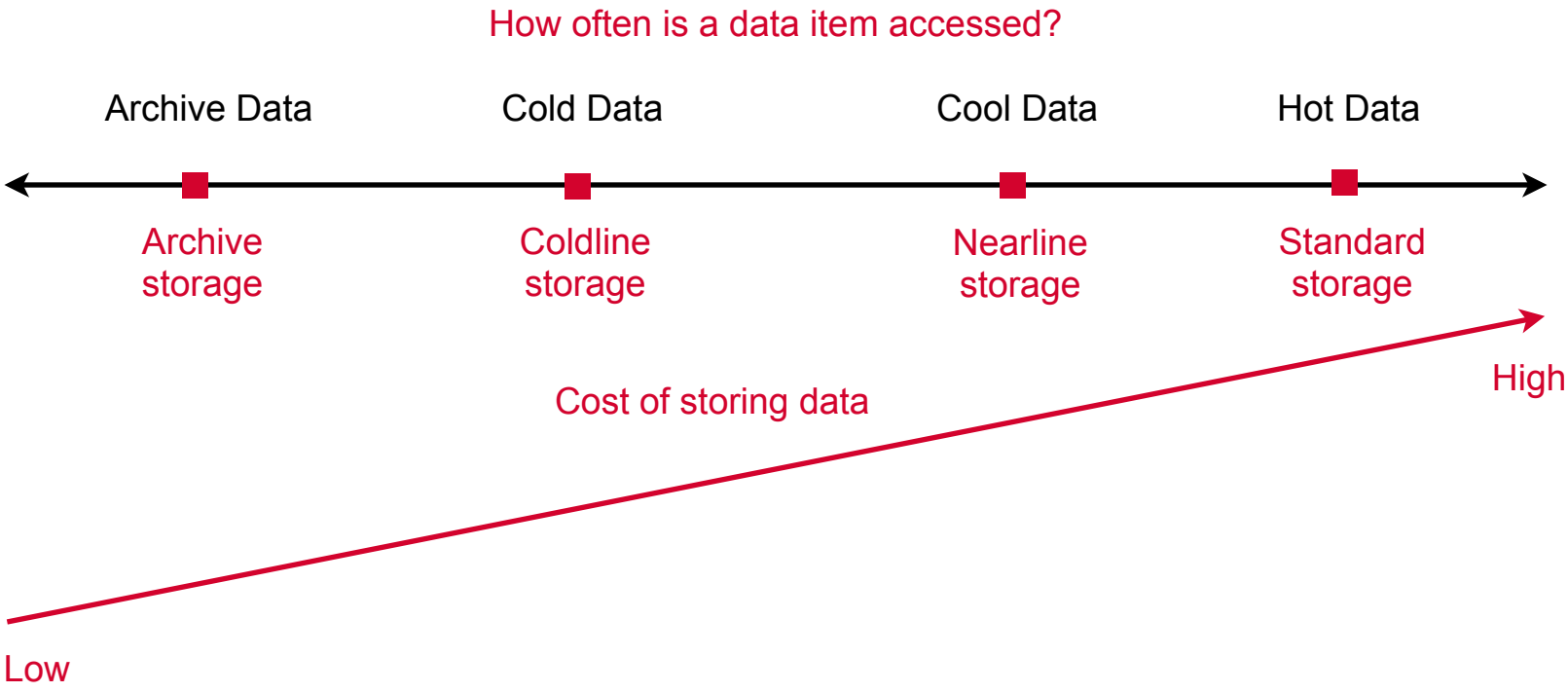
GCS Storage Classes



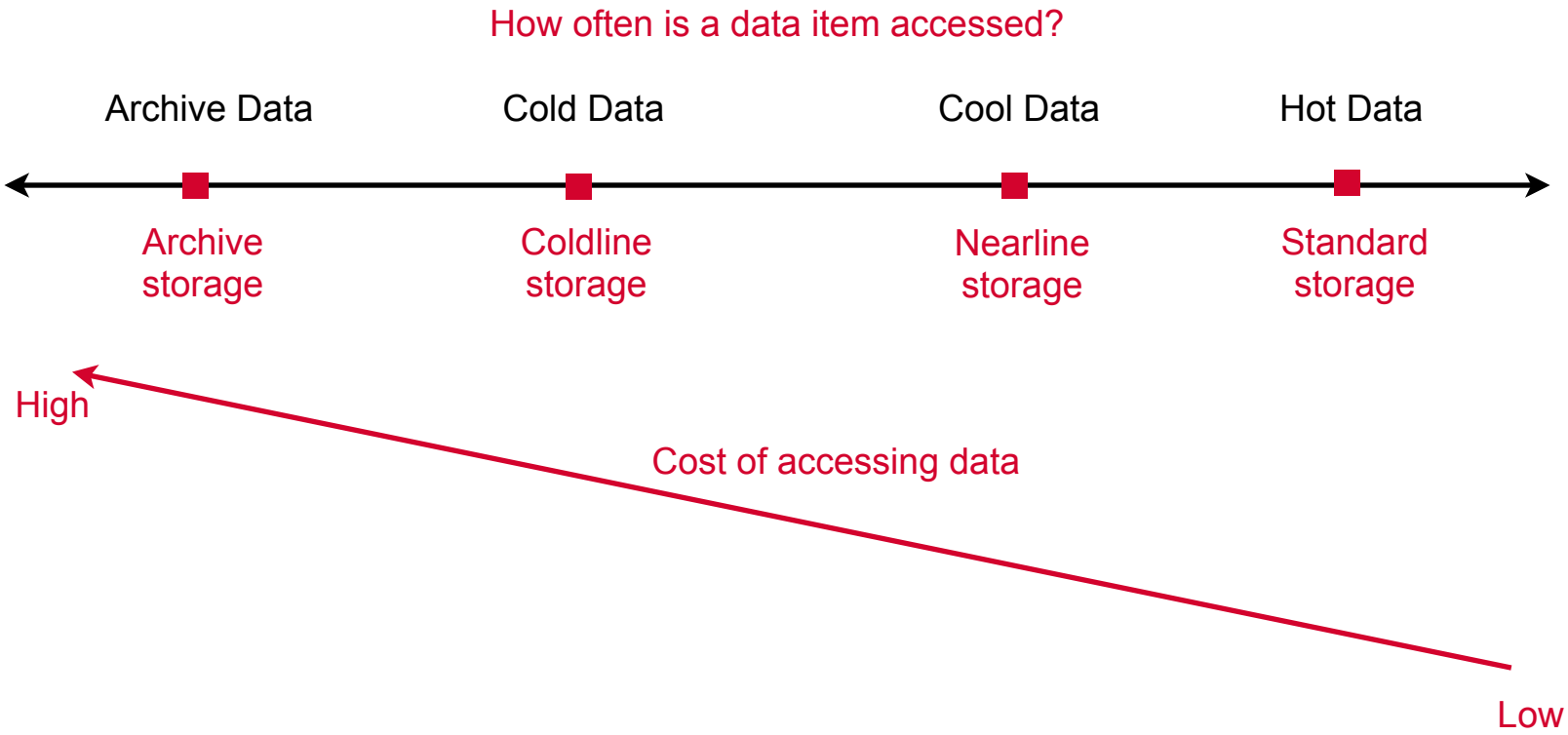
GCS Storage Classes



GCS Storage Classes



GCS Storage Classes



All Storage Classes

Autoclass

Moves data that is not accessed to colder storage classes to reduce cost

Moves data that is accessed to standard storage to optimize cost of future access

Coldline and Archive has about the same speed of access as other storage classes (different from AWS Glacier and S3)

Object Versioning

- Needs to be enabled for bucket
- Once enabled, bucket creates archived versions of each object
- Whenever live object is overwritten or deleted
- Version with unique generation number is created
- Each copy charged separately

Make sure all copies of your data are saved - data is never overwritten

Object Lifecycle Management

- Can automatically specify changes to object storage class
  - “Change from regional to nearline after 30 days”
  - “Delete all data created before 1/8/2018”
  - “Delete all but 2 most recent versions”

Make sure data is available for audit at lower cost

Encryption

- Encrypted even at rest
- Default: Google generates keys
- Can use CSEK (if you want to provide keys)
  - *C*ustomer *S*upplied *E*ncryption *K*ey

Ensure data is stored securely

Retention Policy and Bucket Lock

- Rule on a bucket that forces every object in a bucket to be kept for a certain period of time
  - Compliance and audit requirements e.g. store data for 3 years
- Can optionally apply a bucket lock to ensure that the retention policy cannot be changed
  - Irreversible! Nobody can now change the retention policy

Transfer Appliance

- **Physical, high-capacity storage device** that enables large-scale data migration from your on-premises environment to Google Cloud
- Securely transfer massive amounts of data without relying on network connectivity.
- Data is **encrypted** at rest using AES-256 encryption before it's written to the appliance.
- Google ships the Transfer Appliance to your location.
- You simply connect it to your network, upload your data using standard protocols (like NFS or SMB), and ship it back to Google.
- **Choose when you have to transfer a large amount of data to Google (50 - 100 TB+)**

Storage Transfer Service

- **Fully managed service** to transfer data to Google Cloud from a variety of sources
  - Amazon S3 buckets
  - Azure Blob Storage
  - On-premise file systems
- **Choose over Transfer appliance for small to medium datasets (< 100TB)**
- Can choose for larger datasets when you have sufficient time to perform the transfer (over days or weeks)


O'REILLY®
Other
Storage
Services

Storage Use Cases

Use Case	Appropriate GCP Service	Non-GCP Equivalents
Block storage	Persistent disks or local SSDs	AWS EBS, Azure Disk
Object/blob storage	Cloud Storage (GCS) buckets	AWS S3, Azure Blob Storage
Relational data - small, regional payloads	Cloud SQL	AWS RDS, Azure SQL Database
Relational data - large, global payloads	Cloud Spanner	Aurora DB
HTML/XML documents with NoSQL access	Firestore	AWS DynamoDB, Azure Cosmos DB
Large, naturally ordered data with NoSQL access	BigTable	
Analytics and complex queries with SQL access	BigQuery	AWS Redshift, Azure Synapse Analytics

Cloud SQL

Cloud SQL is the fully-managed MySQL, PostgreSQL and SQL Server database service on the Google Cloud Platform

Transactional support, ACID support

Easiest migration path for on-prem RDBMS

High availability using failover replicas in different zones

Google Cloud Spanner

A global, horizontally scaling, strongly consistent relational database service built on proprietary technology

Scales horizontally by adding nodes

ACID support at scale

Relatively expensive and Google proprietary

Cloud Firestore

Flexible, scalable, NoSQL database for keeping data in sync across client apps.

Mobile and web server development as a part of GCP's Firebase platform

Realtime listeners and offline support

BigQuery Features

- Serverless: No cluster, no provisioning
- Structured data with fields
- Can ingest streaming data at scale
- Autoscaling
- Automatic high availability
- Simple SQL queries

Can partition data based on date/ time fields and expire partitions when no longer needed

Redis

Very popular in-memory key-value NoSQL database

Memcached

General purpose, distributed, memory-caching system

Cloud Memorystore

Google managed service for Redis and Memcached that offers scaling, high availability and a convenient migration path

Google Cloud Bigtable

NoSQL database technology ideal for very large, sparse datasets with sequential ordering in key column; provides very fast writes as well as reads

Choose Bigtable For

- Time series data: Naturally ordered
- Internet of Things data: Constant stream of writes
- Financial data: Often efficiently represented as time series data
- Large datasets > 1 TB with each row < 10 MB

Storage Services

A film studio is creating a central repository on Google Cloud for its global post-production teams. The first step is to migrate their historical archive of 800 TB of raw 8K video footage from their on-premises storage systems.

Furthermore, their active film sets generate approximately 15 TB of new footage *each day*, which must also be uploaded promptly to the cloud repository for editing. The studio's existing internet connection is a 1 Gbps line.

What data transfer strategy should the studio adopt to handle both the initial archive migration and the ongoing daily uploads efficiently?

A. Use Storage Transfer Service to configure a transfer job from the on-premises storage to Cloud Storage. Run this job continuously over the existing 1 Gbps internet link until both the archive and daily footage are uploaded.

B. For the initial 800 TB archive, use Transfer Appliance to ship the data physically to Google. For the ongoing 15 TB daily uploads, establish a Dedicated Interconnect of at least 10 Gbps.

C. Use Transfer Appliance for the initial 800 TB archive. For the daily uploads, set up multiple Cloud VPN tunnels over the existing 1 Gbps internet connection to aggregate bandwidth.

D. For the initial archive, provision a 10 Gbps Dedicated Interconnect and transfer the 800 TB over this link. For the daily uploads, use Transfer Appliance, ordering a new one each day.

Storage Services

A film studio is creating a central repository on Google Cloud for its global post-production teams. The first step is to migrate their historical archive of 800 TB of raw 8K video footage from their on-premises storage systems.

Furthermore, their active film sets generate approximately 15 TB of new footage *each day*, which must also be uploaded promptly to the cloud repository for editing. The studio's existing internet connection is a 1 Gbps line.

What data transfer strategy should the studio adopt to handle both the initial archive migration and the ongoing daily uploads efficiently?

A. Use Storage Transfer Service to configure a transfer job from the on-premises storage to Cloud Storage. Run this job continuously over the existing 1 Gbps internet link until both the archive and daily footage are uploaded.

B. For the initial 800 TB archive, use Transfer Appliance to ship the data physically to Google. For the ongoing 15 TB daily uploads, establish a Dedicated Interconnect of at least 10 Gbps.

C. Use Transfer Appliance for the initial 800 TB archive. For the daily uploads, set up multiple Cloud VPN tunnels over the existing 1 Gbps internet connection to aggregate bandwidth.

D. For the initial archive, provision a 10 Gbps Dedicated Interconnect and transfer the 800 TB over this link. For the daily uploads, use Transfer Appliance, ordering a new one each day.

Storage Services

A logistics company is building a data lake on Google Cloud to ingest real-time JSON telemetry data from thousands of IoT sensors on its trucks. The structure of the JSON payload can change unexpectedly when sensor firmware is updated.

A key business requirement is that the raw, unaltered JSON messages from every sensor must be preserved indefinitely in their original format. This ensures that data can be reprocessed if a downstream analytics pipeline fails due to a schema change.

What is the most appropriate initial step for the data ingestion architecture to meet this requirement?

A. Stream the JSON data directly into a Dataflow pipeline that validates and transforms the data. Store the resulting structured data in BigQuery.

B. Create a Cloud Storage bucket to serve as the landing zone. Ingest the raw JSON messages and store each one as an individual, immutable object in this bucket before any processing occurs.

C. Ingest the JSON messages directly into a Cloud Bigtable table, with each message's fields stored in separate columns. Design the processing pipelines to read from Bigtable.

D. Publish all incoming JSON messages to a Pub/Sub topic. Set the message retention duration on the topic to its maximum period to allow for reprocessing from the topic itself.

Storage Services

A logistics company is building a data lake on Google Cloud to ingest real-time JSON telemetry data from thousands of IoT sensors on its trucks. The structure of the JSON payload can change unexpectedly when sensor firmware is updated.

A key business requirement is that the raw, unaltered JSON messages from every sensor must be preserved indefinitely in their original format. This ensures that data can be reprocessed if a downstream analytics pipeline fails due to a schema change.

What is the most appropriate initial step for the data ingestion architecture to meet this requirement?

A. Stream the JSON data directly into a Dataflow pipeline that validates and transforms the data. Store the resulting structured data in BigQuery.

B. Create a Cloud Storage bucket to serve as the landing zone. Ingest the raw JSON messages and store each one as an individual, immutable object in this bucket before any processing occurs.

C. Ingest the JSON messages directly into a Cloud Bigtable table, with each message's fields stored in separate columns. Design the processing pipelines to read from Bigtable.

D. Publish all incoming JSON messages to a Pub/Sub topic. Set the message retention duration on the topic to its maximum period to allow for reprocessing from the topic itself.

Storage Solutions

You are designing a backend service for an e-commerce platform that will handle and store transactional data from mobile and web users located worldwide. With the platform's global launch, you expect to handle petabytes of transaction data, and the business team needs the ability to run SQL queries for analysis. You need to create a data store that is highly available, scalable, and capable of efficiently handling massive volumes of data. What should you do?

A. Use Cloud Spanner to create a globally distributed, scalable, and highly available relational database that supports SQL queries.
B. Use Firestore in Datastore mode to store transactional data and run SQL queries through third-party tools.
C. Set up a Cloud SQL instance with regional replication to handle the large volume of global transactions.
D. Deploy a MySQL database on Compute Engine VMs and configure replication across regions to handle the data load.

Storage Solutions

You are designing a backend service for an e-commerce platform that will handle and store transactional data from mobile and web users located worldwide. With the platform's global launch, you expect to handle petabytes of transaction data, and the business team needs the ability to run SQL queries for analysis. You need to create a data store that is highly available, scalable, and capable of efficiently handling massive volumes of data. What should you do?

A. **Use Cloud Spanner to create a globally distributed, scalable, and highly available relational database that supports SQL queries.**
B. Use Firestore in Datastore mode to store transactional data and run SQL queries through third-party tools.
C. Set up a Cloud SQL instance with regional replication to handle the large volume of global transactions.
D. Deploy a MySQL database on Compute Engine VMs and configure replication across regions to handle the data load.

Storage Solutions

You are designing a system to ingest and store data from millions of IoT sensors around the world. The data will be sent continuously at high throughput, and the business team needs to query and analyze this data based on the time it was generated. You want to ensure that the system can scale efficiently and handle real-time queries. What should you do?

A. Ingest the data into BigQuery, and partition the tables by timestamp for querying.
B. Ingest the data into Cloud SQL and index it by timestamp for efficient queries.
C. Ingest the data into Bigtable, and create a row key based on the event timestamp for optimized querying and scaling.
D. Ingest the data into Firestore, and organize the data based on event timestamps for real-time analysis.

Storage Solutions

You are designing a system to ingest and store data from millions of IoT sensors around the world. The data will be sent continuously at high throughput, and the business team needs to query and analyze this data based on the time it was generated. You want to ensure that the system can scale efficiently and handle real-time queries. What should you do?

A. Ingest the data into BigQuery, and partition the tables by timestamp for querying.
B. Ingest the data into Cloud SQL and index it by timestamp for efficient queries.
C. Ingest the data into Bigtable, and create a row key based on the event timestamp for optimized querying and scaling.
D. Ingest the data into Firestore, and organize the data based on event timestamps for real-time analysis.

Storage Solutions

Your company is deploying a business-critical application that requires a relational database on Google Cloud. The application must be highly available and resilient to zonal failures. You need to ensure that the database remains operational even if an entire zone becomes unavailable. What should you do to meet these requirements?

A. Use a single Cloud SQL instance and configure automated backups for disaster recovery.
B. Use Cloud SQL configured for high availability with failover replicas in different zones.
C. Deploy a Cloud SQL instance in a single zone and set up replication using Cloud Spanner.
D. Use Bigtable with replication in the same zone to ensure availability and performance.

Storage Solutions

Your company is deploying a business-critical application that requires a relational database on Google Cloud. The application must be highly available and resilient to zonal failures. You need to ensure that the database remains operational even if an entire zone becomes unavailable. What should you do to meet these requirements?

A. Use a single Cloud SQL instance and configure automated backups for disaster recovery.

B. Use Cloud SQL configured for high availability with failover replicas in different zones.

C. Deploy a Cloud SQL instance in a single zone and set up replication using Cloud Spanner.

D. Use Bigtable with replication in the same zone to ensure availability and performance.

O'REILLY®

Identity and Access Management

Cloud IAM

Manage identity and access control by defining who (identity) has what access (role) for which resource.

Cloud IAM

Permission to access a resource is not granted directly to the end user. Instead, permissions are grouped into roles, and roles are granted to authenticated principals.

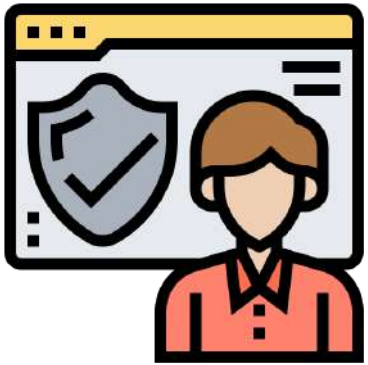
Cloud IAM

Permission to access a resource is not granted directly to the end user. Instead, permissions are grouped into roles, and roles are granted to authenticated principals.

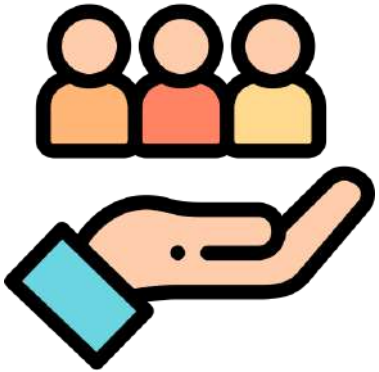
- Principal: GCP identity - user, group, service account
- Role: Collection of permissions
- Policy: Binding members to a role

Role-based Access Control

Identity



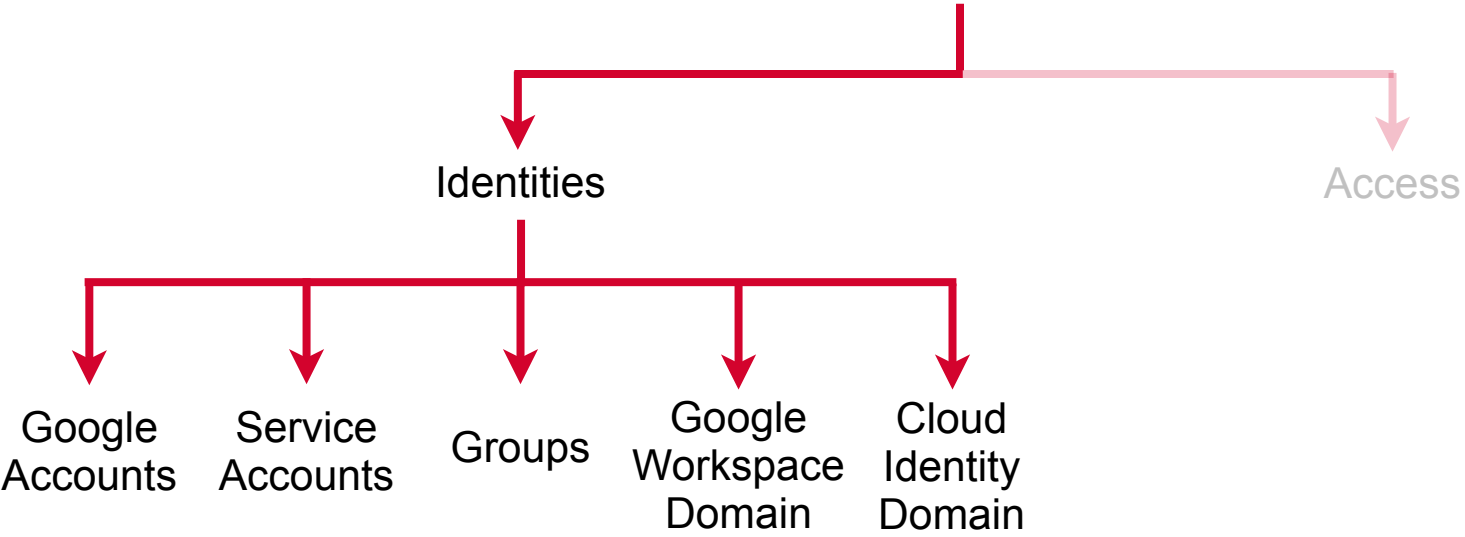
Permissions



Resource

Identity and Access Management (IAM)

Identity and Access Management (IAM)



Google Accounts

A Google account represents a developer, an administrator, or any other person who interacts with GCP.

Service Accounts

A service account is an account that belongs to your application instead of to an individual end user.

Google Groups

A Google Group is a named collection of Google accounts and service accounts. Every group has a unique email address that is associated with the group.

Google Workspace Domains

A Google Workspace domain represents a virtual group of all the Google accounts that have been created in an organization’s account.

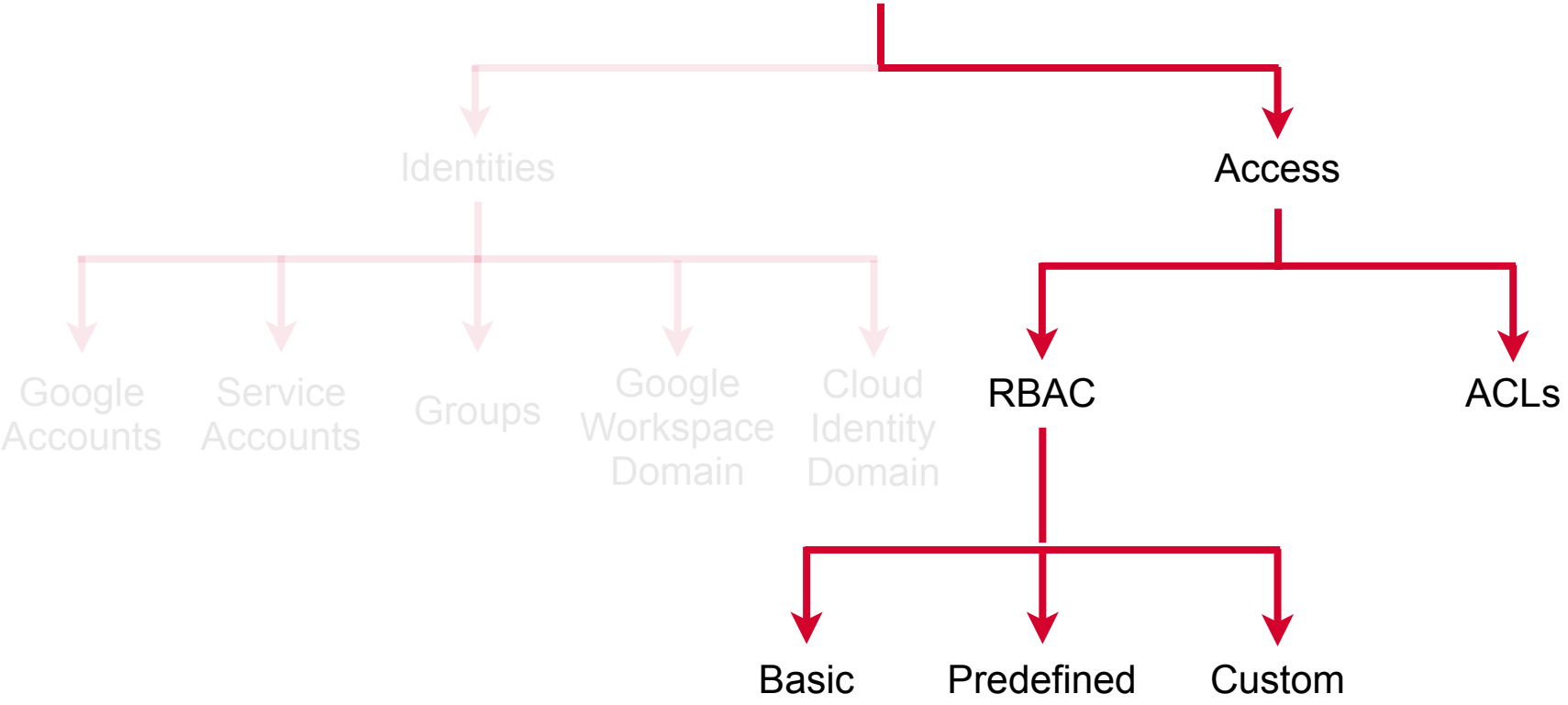
Google Workspace domains represent your organization's Internet domain name.

Cloud Identity Domains

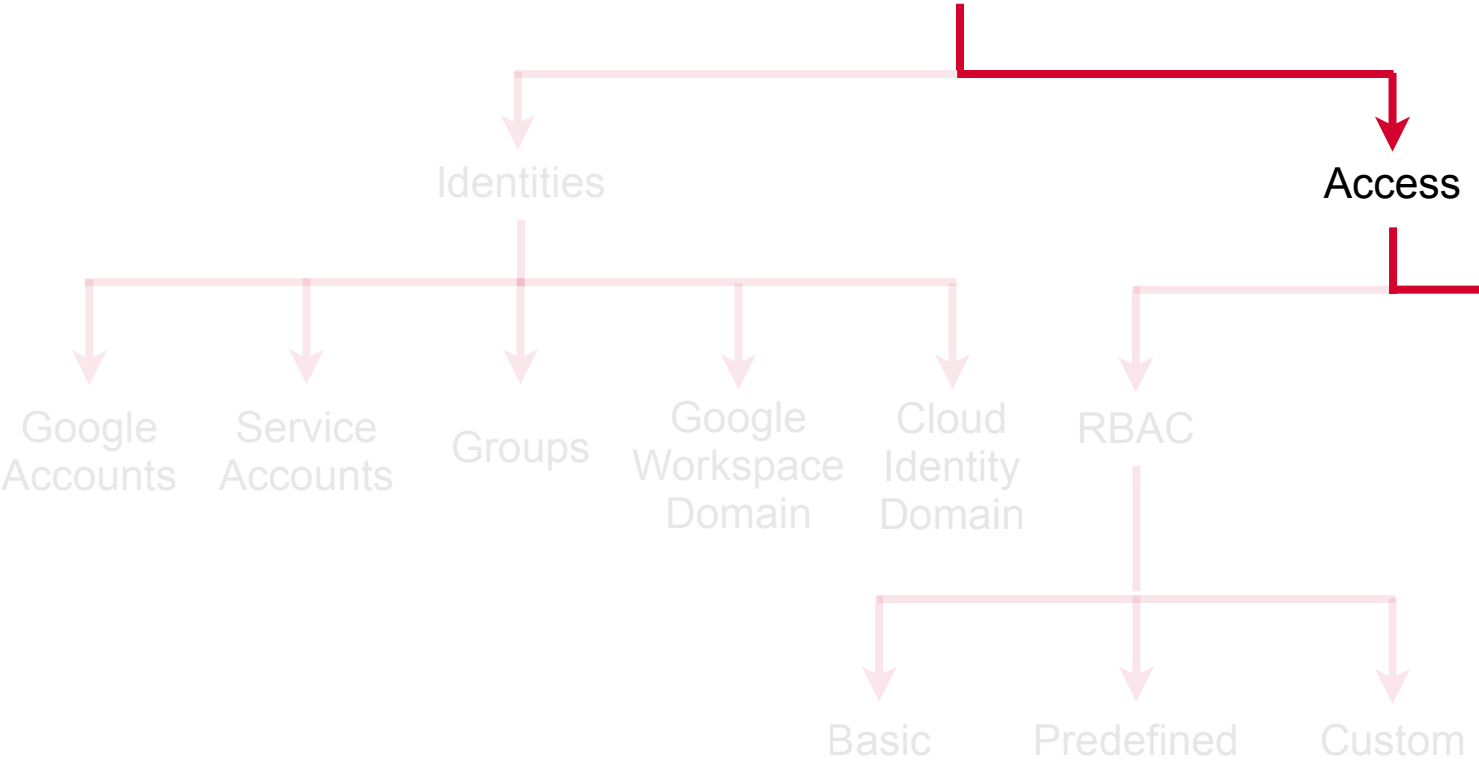
A Cloud Identity domain is like a Google Workspace domain because it represents a virtual group of all Google accounts in an organization.

However, Cloud Identity domain users don't have access to Google Workspace applications and features.

Identity and Access Management (IAM)



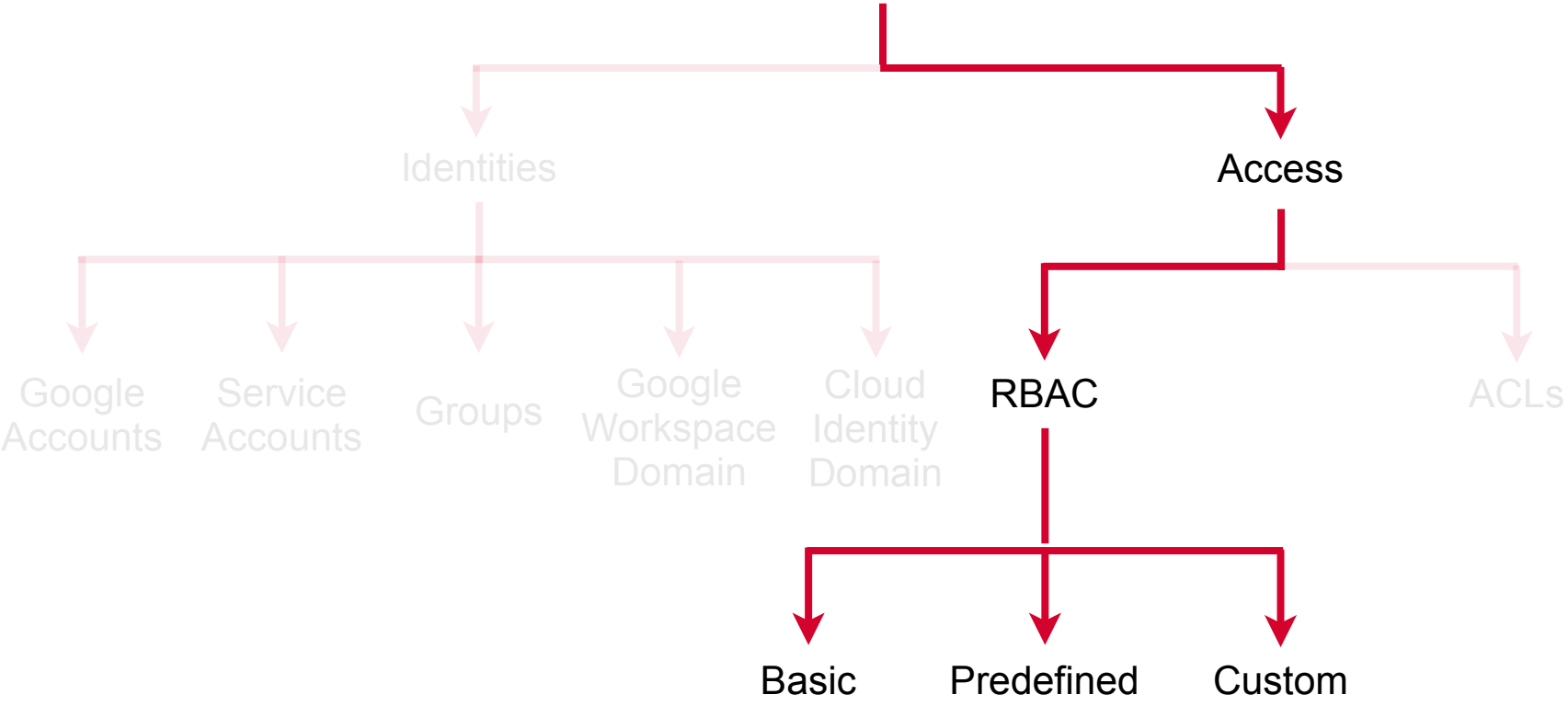
ACLs Not Part of the IAM Service on Google



ACLs

Directly specify which users or service accounts have access to resources. Only supported by **Google Cloud Storage for PII**

Identity and Access Management (IAM)



Basic Roles

Three concentric roles that existed prior to the introduction of Cloud IAM: Owner, Editor, and Viewer of any resource.

Historically available, not recommended unless there is not alternative.

Predefined Roles

- Project Roles
- App Engine Roles
- BigQuery Roles
- Cloud Bigtable Roles
- Cloud Billing Roles

Predefined Roles

roles/bigquery.dataViewer

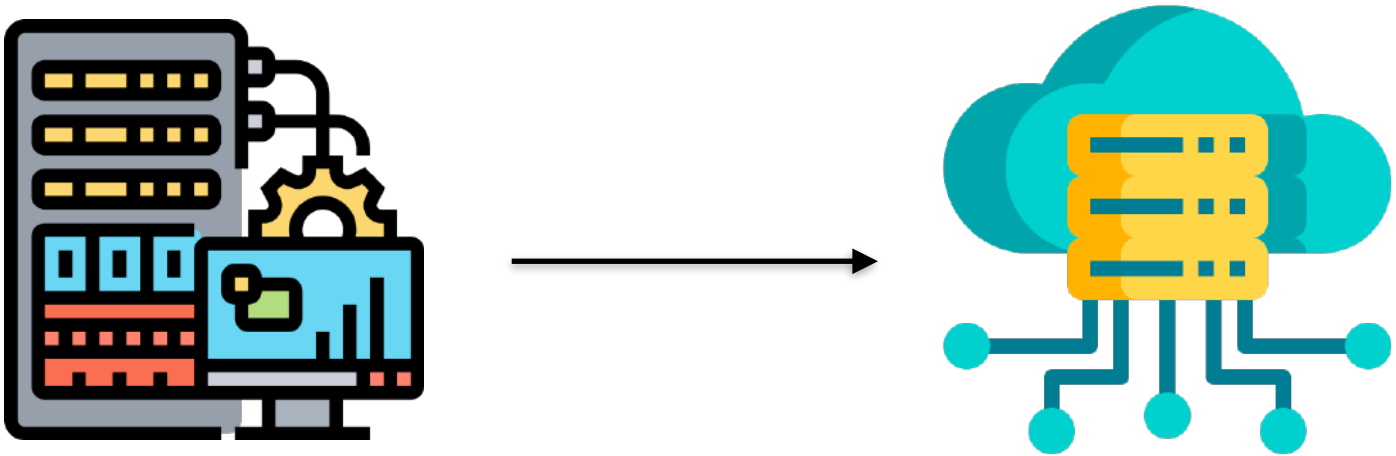
bigquery.datasets.get
bigquery.datasets.getIamPolicy
bigquery.models.getData
bigquery.models.getMetadata
bigquery.models.list
bigquery.routines.get
bigquery.routines.list
bigquery.tables.export
bigquery.tables.get
bigquery.tables.getData
bigquery.tables.list
resourcemanager.projects.get
resourcemanager.projects.list

Custom Roles

User-defined roles that bundle one or more supported permissions tailored to meet your specific needs.

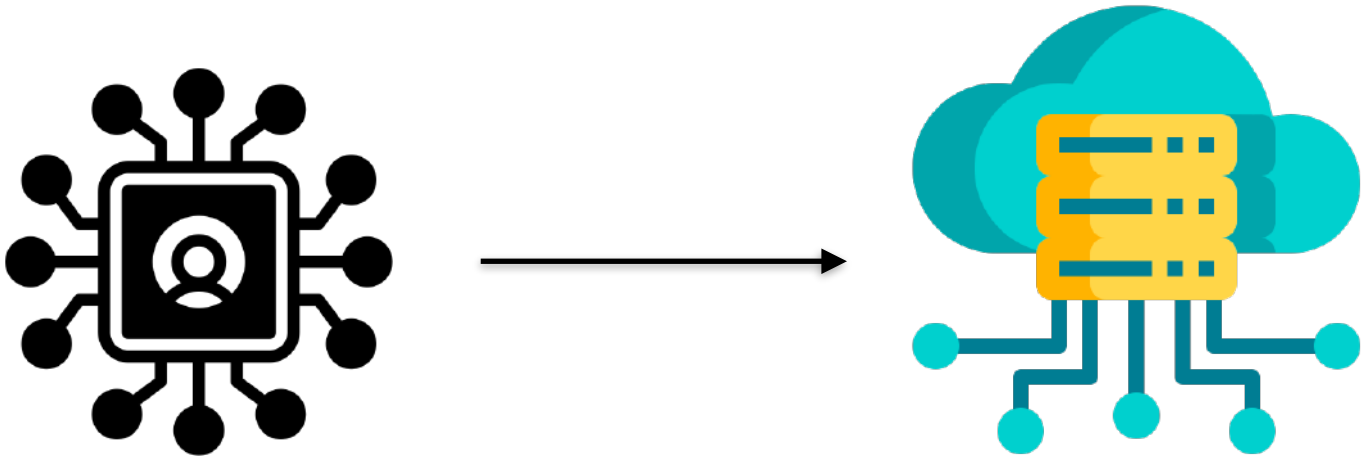
Not maintained by Google; when new permissions, features, or services are added to GCP, your custom roles will not be updated automatically.

Applications Accessing Cloud Resources



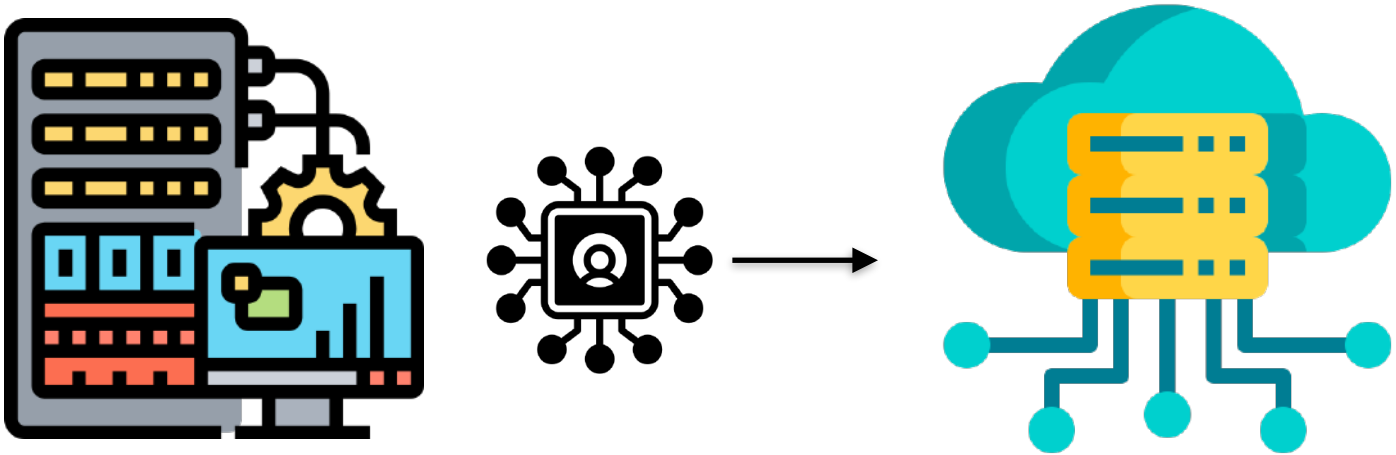
Your application running on the server needs to access cloud resources as a part of its execution - how do you assign permissions to your application?

Service Accounts



Create a service account and give it access
to the resource on the cloud

Configure Application to Use the Service Account

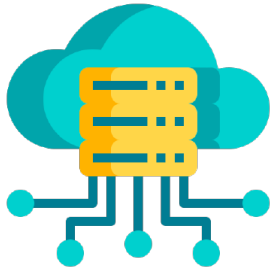


The application will run under the identity of the service account and use the roles granted to the service account to access resources on the cloud

Applications running in one project can access services running in another project using service accounts

Accessing Cloud Resources Across Projects

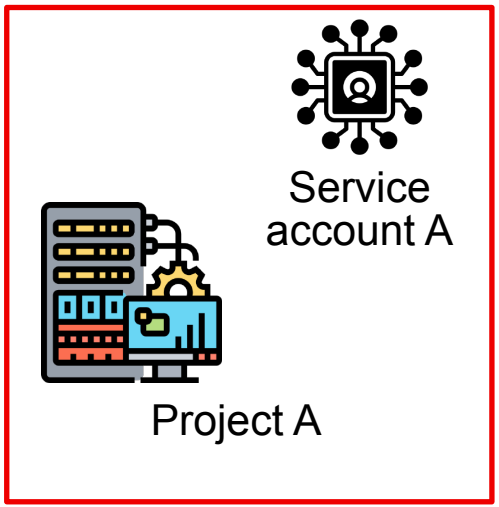
Project A

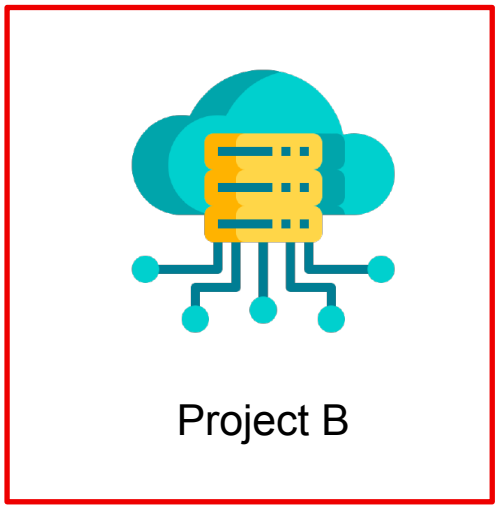


Project B

Application running in project A needs access
to a database in project B

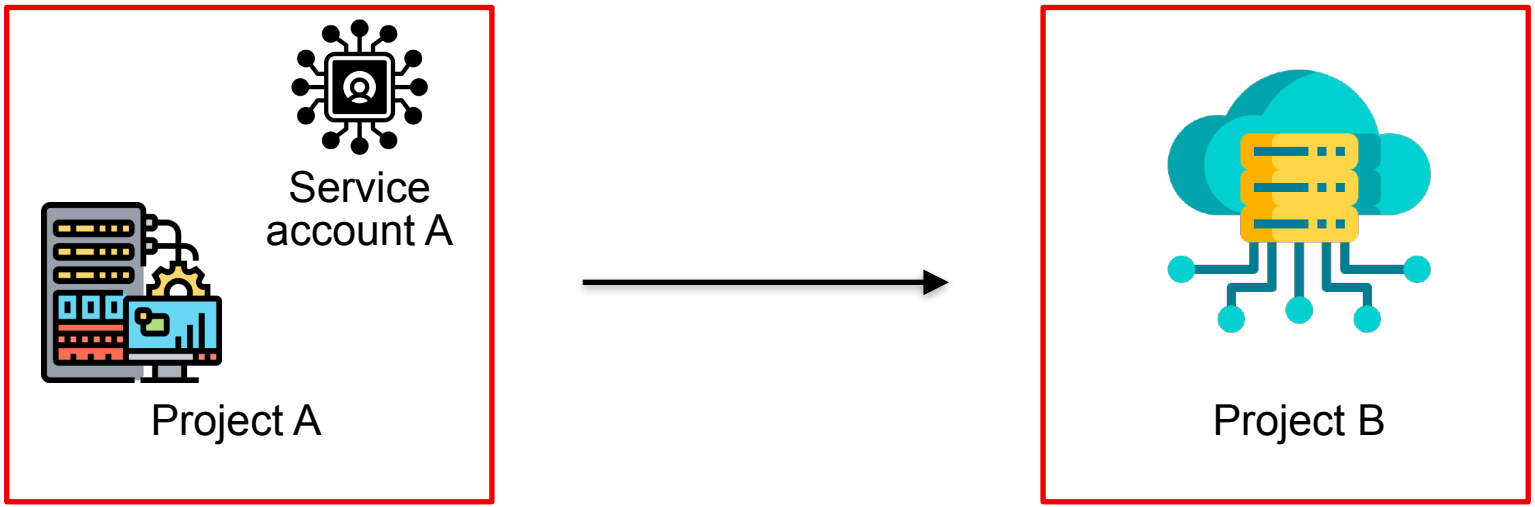
Create a Service Account in Project A





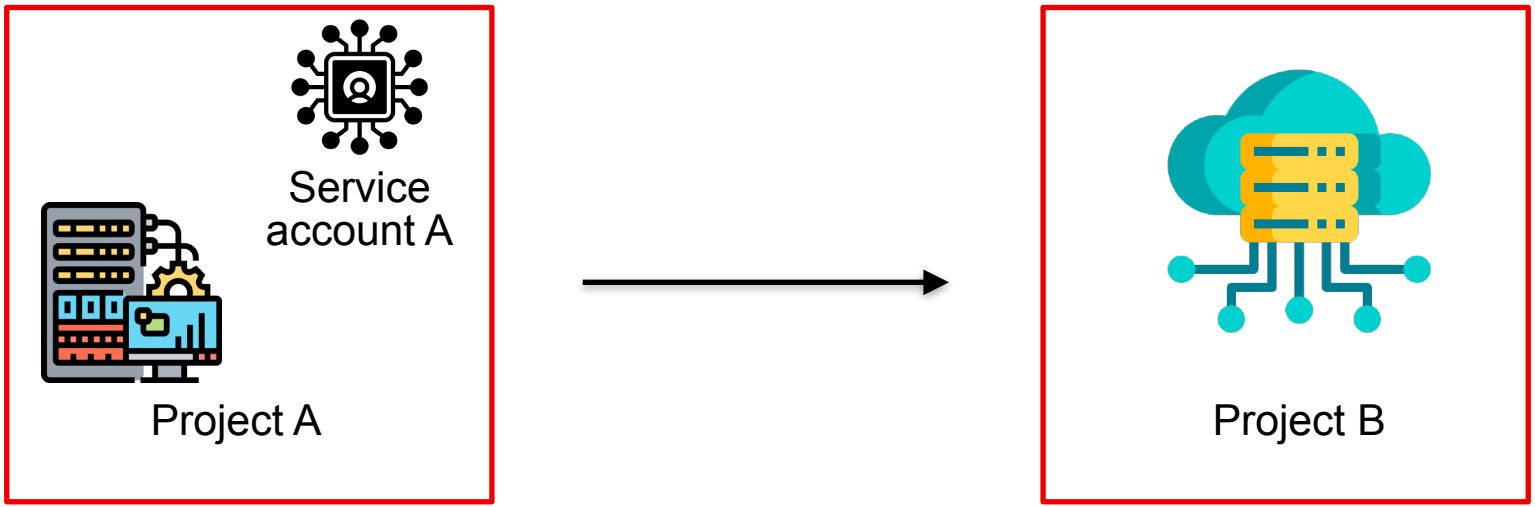
Project A owners create a service account in Project A

Configure Service Account to Access Resources



Project B owners give service account A access to the database in Project B

Accessing Cloud Resources Across Projects



Applications running in Project A can now access the database in Project B


O'REILLY®
Identity
Aware
Proxy

Identity Aware Proxy (IAP)

A central authorization layer for applications accessed by HTTPS, so you can use an application-level access control model instead of relying on network-level firewalls.

Define access policies centrally and apply them to all of your applications and resources.

Can set up individual or group-based access to applications

IAM and IAP (Identity Aware Proxy)

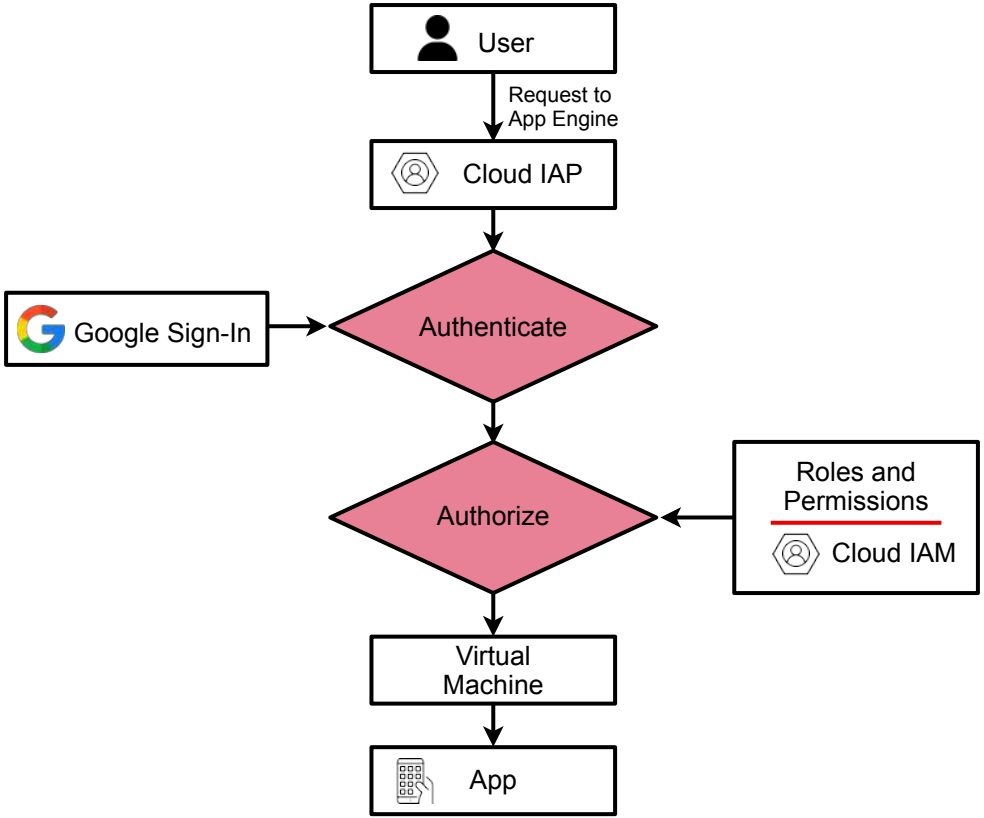
IAM

- Access controls and permissions for Google Cloud resources
- Configured at the resource level (VMs, buckets, datasets)

IAP

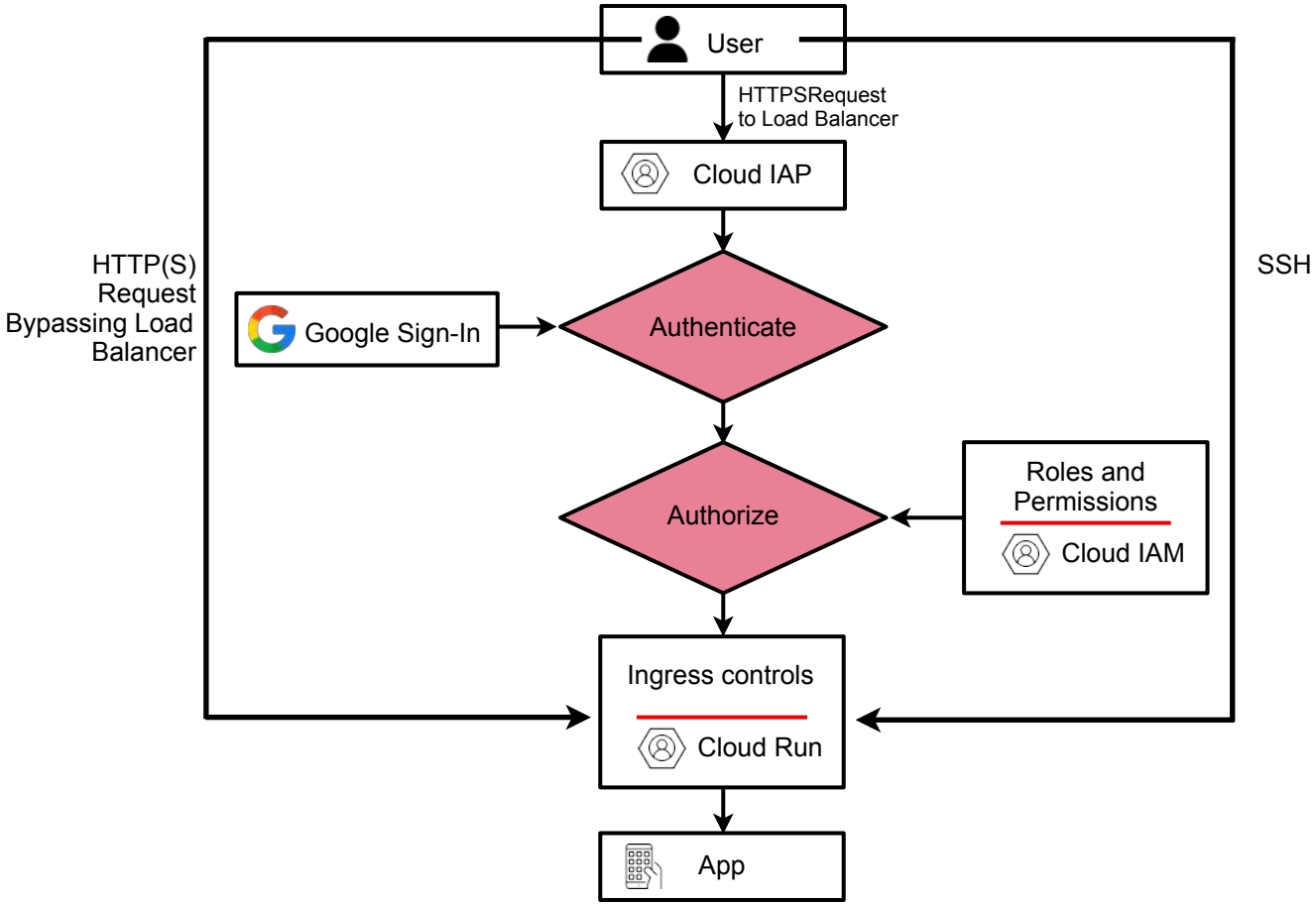
- Security layer that controls access to applications running on Google Cloud
- Configured to protect applications by intercepting requests to them
- Uses identities and roles from IAM to grant access to applications

IAP with App Engine



Can work with Cloud Run, Compute Engine, GKE, and even On-premise apps

IAP with App Engine



IAP secures authentication and authorization of all requests to App Engine, Cloud Load Balancing (HTTPS), or internal HTTP load balancing.

IAP doesn't protect against activity within a project, such as another VM inside the project.

IAM

A financial corporation uses two separate Google Cloud projects: finance-data-project for storing sensitive quarterly reports in a Cloud Storage bucket, and analytics-pipeline-project for running a data processing pipeline on Compute Engine.

The pipeline running in analytics-pipeline-project needs programmatic, read-only access to the reports in the Cloud Storage bucket located in finance-data-project. You must provide this access following the principle of least privilege, granting only the absolute minimum permissions required.

What is the most secure and appropriate method to grant this access?

A. Move the Compute Engine instance from the analytics-pipeline-project into the finance-data-project so they are in the same project and can communicate directly.

B. In the finance-data-project, grant the Project Editor role to the default Compute Engine service account from the analytics-pipeline-project.

C. In the finance-data-project, create a new **service account**. Grant this service account the **Storage Object Viewer** role on the specific bucket. Then, grant the Compute Engine service account in the analytics-pipeline-project the permission to impersonate this new service account.

D. Make the Cloud Storage bucket public so the analytics pipeline can read the data, but use a firewall rule to restrict access to only the IP address of the Compute Engine instance.

IAM

A financial corporation uses two separate Google Cloud projects: finance-data-project for storing sensitive quarterly reports in a Cloud Storage bucket, and analytics-pipeline-project for running a data processing pipeline on Compute Engine.

The pipeline running in analytics-pipeline-project needs programmatic, read-only access to the reports in the Cloud Storage bucket located in finance-data-project. You must provide this access following the principle of least privilege, granting only the absolute minimum permissions required.

What is the most secure and appropriate method to grant this access?

A. Move the Compute Engine instance from the analytics-pipeline-project into the finance-data-project so they are in the same project and can communicate directly.

B. In the finance-data-project, grant the Project Editor role to the default Compute Engine service account from the analytics-pipeline-project.

C. In the finance-data-project, create a new service account. Grant this service account the Storage Object Viewer role on the specific bucket. Then, grant the Compute Engine service account in the analytics-pipeline-project the permission to impersonate this new service account.

D. Make the Cloud Storage bucket public so the analytics pipeline can read the data, but use a firewall rule to restrict access to only the IP address of the Compute Engine instance.

IAM

A European bank is migrating to Google Cloud and must adhere to strict GDPR data sovereignty rules. Their corporate policy dictates that all cloud resources and stored data must physically reside within European Union data centers.

What is the most effective and scalable Google Cloud feature to enforce this policy?

A. Submit requests to Google Cloud Support to reduce the resource quotas for every service to zero in all non-EU regions.

B. Define an Organization Policy with the gcp.resourceLocations constraint. In the policy, specify the list of allowed Google Cloud locations to include only the EU regions.

C. Create a Cloud Audit Logs sink that exports all resource creation logs to BigQuery. Set up a scheduled query to run every hour to report on any resources created outside the EU. Set up a script to immediately shutdown those resources

D. Configure the organization's VPC firewall rules to deny all traffic that originates from or is destined for IP ranges associated with non-EU regions.

IAM

A European bank is migrating to Google Cloud and must adhere to strict GDPR data sovereignty rules. Their corporate policy dictates that all cloud resources and stored data must physically reside within European Union data centers.

What is the most effective and scalable Google Cloud feature to enforce this policy?

A. Submit requests to Google Cloud Support to reduce the resource quotas for every service to zero in all non-EU regions.

B. Define an Organization Policy with the gcp.resourceLocations constraint. In the policy, specify the list of allowed Google Cloud locations to include only the EU regions.

C. Create a Cloud Audit Logs sink that exports all resource creation logs to BigQuery. Set up a scheduled query to run every hour to report on any resources created outside the EU. Set up a script to immediately shutdown those resources

D. Configure the organization's VPC firewall rules to deny all traffic that originates from or is destined for IP ranges associated with non-EU regions.

IAM

You are deploying a new virtual machine (VM) on Google Cloud to run a specific batch processing job. The VM needs to read data from a Cloud Storage bucket and write results to a Cloud SQL instance. You want to configure its access to these services by following the principle of least privilege and Google Cloud's recommended security practices.

What should you do?

A. Modify the project's default Compute Engine service account by adding the Cloud SQL Client and Storage Object Admin roles to it. Attach this default service account to the new VM.

B. Create a new, dedicated service account for the VM. To ensure it has sufficient permissions, grant it the project-level Editor role. Attach this service account to the VM.

C. Create a new, dedicated service account and grant it the necessary roles (Cloud SQL Client, Storage Object Viewer). Generate a JSON key for this service account, securely copy it to the VM, and configure the application to use this key for authentication.

D. Create a new, dedicated service account and grant it the specific, predefined roles it needs (Cloud SQL Client, Storage Object Viewer). Attach this service account to the VM during its creation.

IAM

You are deploying a new virtual machine (VM) on Google Cloud to run a specific batch processing job. The VM needs to read data from a Cloud Storage bucket and write results to a Cloud SQL instance. You want to configure its access to these services by following the principle of least privilege and Google Cloud's recommended security practices.

What should you do?

A. Modify the project's default Compute Engine service account by adding the Cloud SQL Client and Storage Object Admin roles to it. Attach this default service account to the new VM.

B. Create a new, dedicated service account for the VM. To ensure it has sufficient permissions, grant it the project-level Editor role. Attach this service account to the VM.

C. Create a new, dedicated service account and grant it the necessary roles (Cloud SQL Client, Storage Object Viewer). Generate a JSON key for this service account, securely copy it to the VM, and configure the application to use this key for authentication.

D. Create a new, dedicated service account and grant it the specific, predefined roles it needs (Cloud SQL Client, Storage Object Viewer). Attach this service account to the VM during its creation.

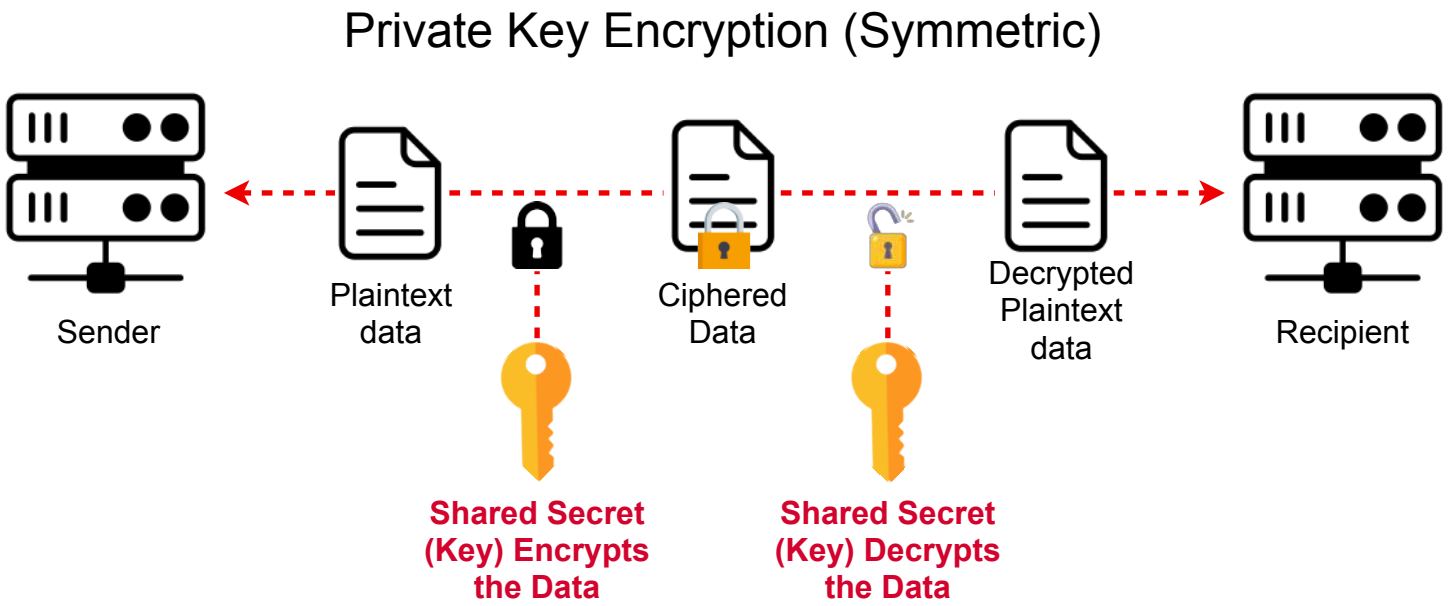
Key Management

Cryptographic Keys

Cryptographic keys serve as the secret codes that enable the encryption and decryption of data.

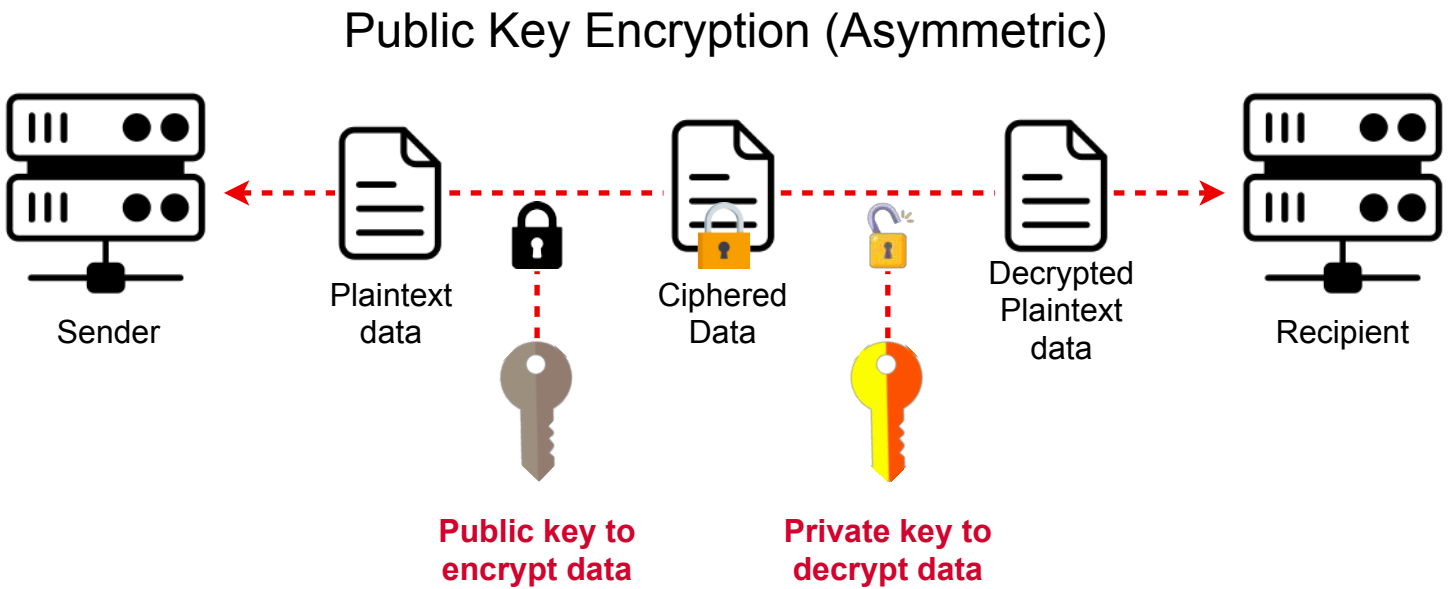
Ensure that only authorized parties with the correct key can access and read the encrypted information

Symmetric Key Encryption



Encrypting the data and decrypting the data make use of the same shared key

Asymmetric Key Encryption



The encryption key is publicly available - the decryption key is private

Default Encryption on the Google Cloud

All Google Cloud services that store data encrypt data by default

No configuration and automatic encryption. Most services automatically rotate keys

Google-owned and Google-managed keys

Customer Managed Encryption Keys (CMEK)

Encryption keys that customers create, own, and manage within cloud services to secure their data

CMEKs give customers greater control over their encryption practices, including key rotation and access policies

Cloud Key Management Service (CMEKs)

Software-based keys

Hardware-based keys

External keys

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Cloud HSM (Hardware Security Module)

Hardware-based keys

Cloud EKM (External Key Manager)

External keys

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Cloud HSM (Hardware Security Module)

Hardware-based keys

Cloud EKM (External Key Manager)

External keys

Control keys, key rotation schedule, IAM roles and permissions

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Control keys, key rotation schedule, IAM roles and permissions

Cloud HSM (Hardware Security Module)

Hardware-based keys

More secure that software keys, stored in a separate physical device

Control keys, key rotation schedule, IAM roles and permissions

Cloud EKM (External Key Manager)

External keys

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Control keys, key rotation schedule, IAM roles and permissions

Cloud HSM (Hardware Security Module)

Hardware-based keys

More secure that software keys, stored in a separate physical device

Control keys, key rotation schedule, IAM roles and permissions

Cloud EKM (External Key Manager)

External keys

Keys stored outside Google in an external provider, keys never sent to Google

Control keys, key rotation schedule, IAM roles and permissions

Customer Supplied Encryption Keys (CSEK)

Customers provide key materials when needed.

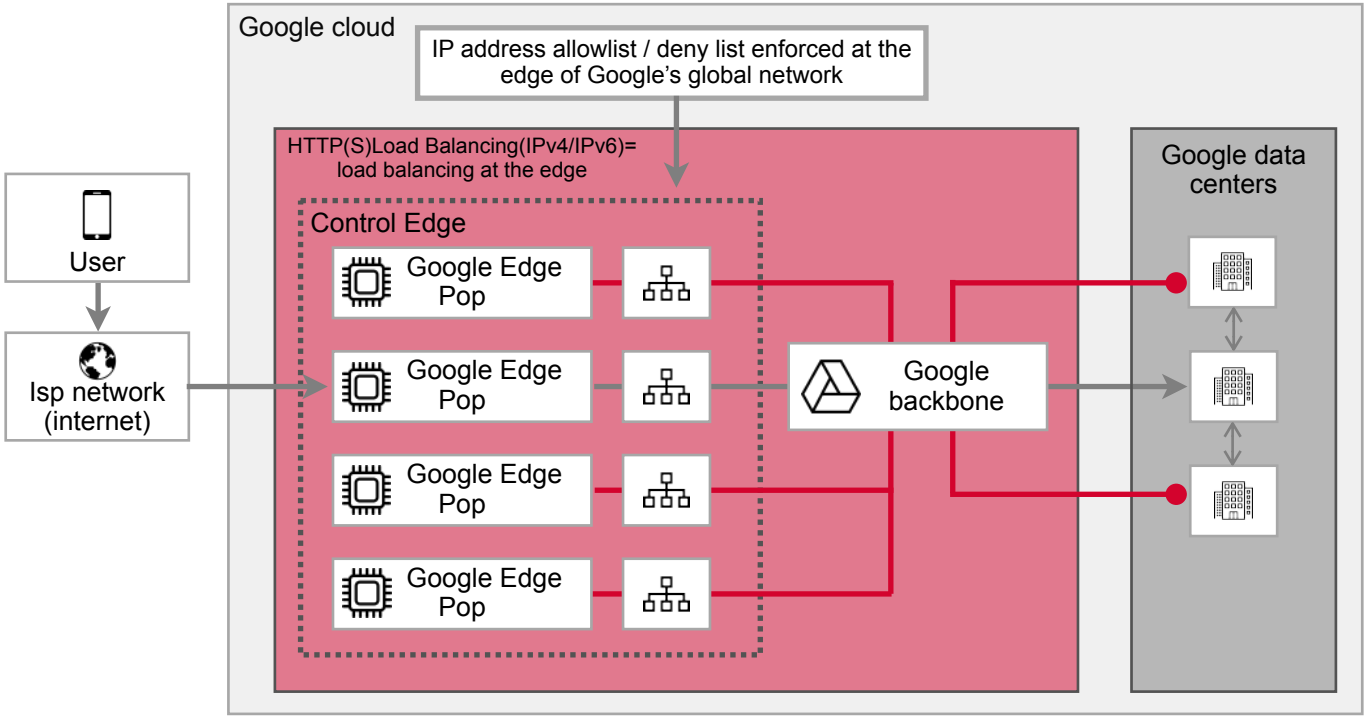
Google keeps keys in-memory, keys not stored permanently on Google's servers


O'REILLY®
Cloud Armor and Data Loss Prevention

Cloud Armor

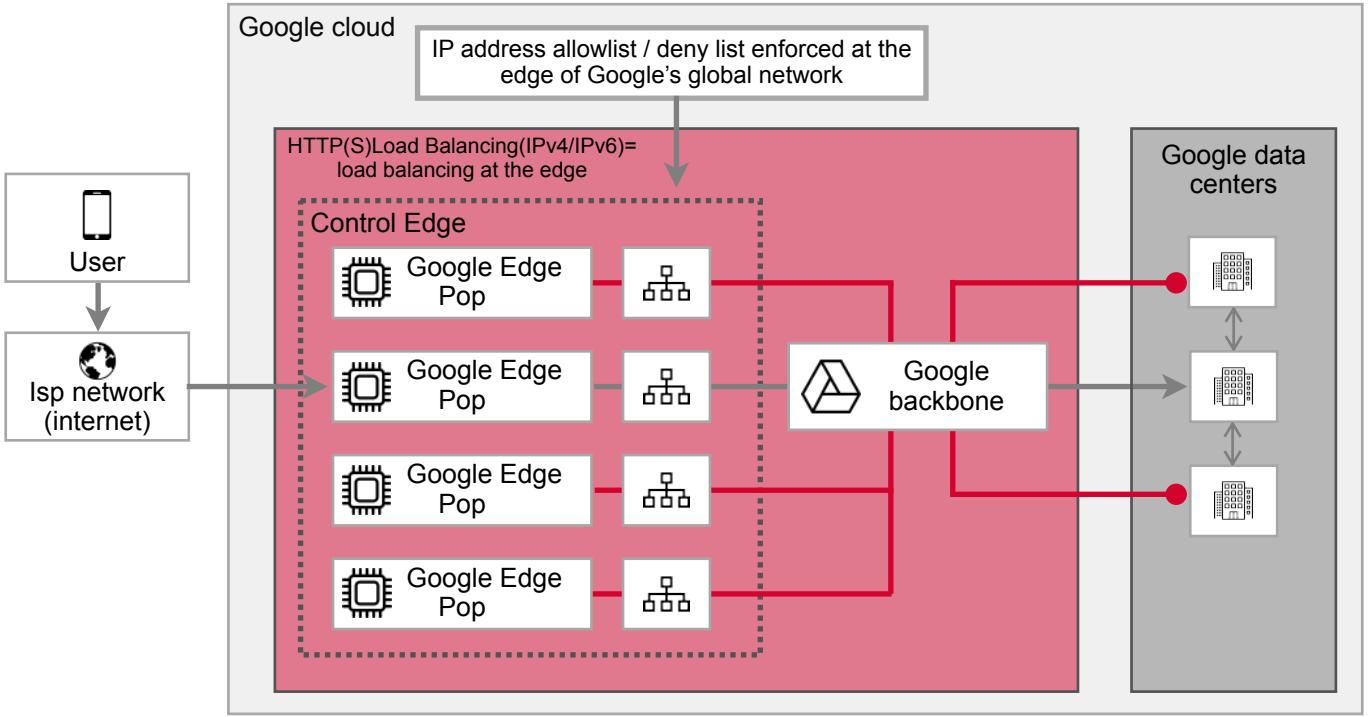
Helps protect your Google Cloud deployments from multiple types of threats, including distributed denial-of-service (DDoS) attacks, cross-site scripting (XSS), and SQL injection (SQLi).

How Does Cloud Armor Work?



Protection against volumetric DDoS attacks. Protection for applications and services running behind a load balancer

How Does Cloud Armor Work?



Security policies enforce custom Layer 7 filtering policies including preconfigured web application firewall (WAF) rules to mitigate OWASP top 10 web application vulnerability risks

Cloud Armor protection is for external requests coming into your Load Balancer

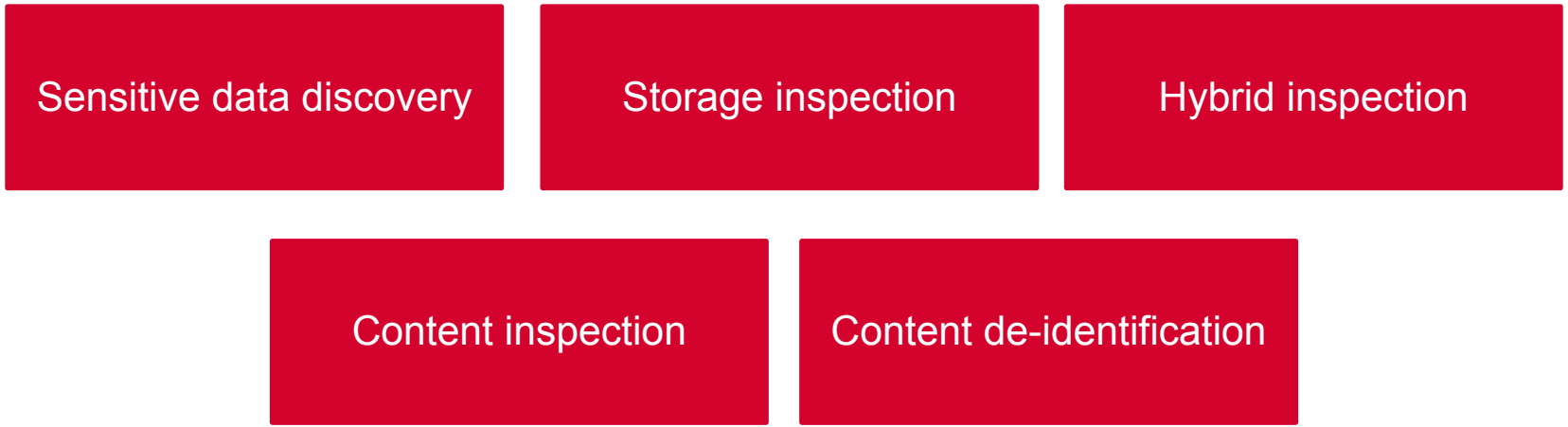
Firewall rules govern traffic flows inside your VPC

Sensitive Data Protection

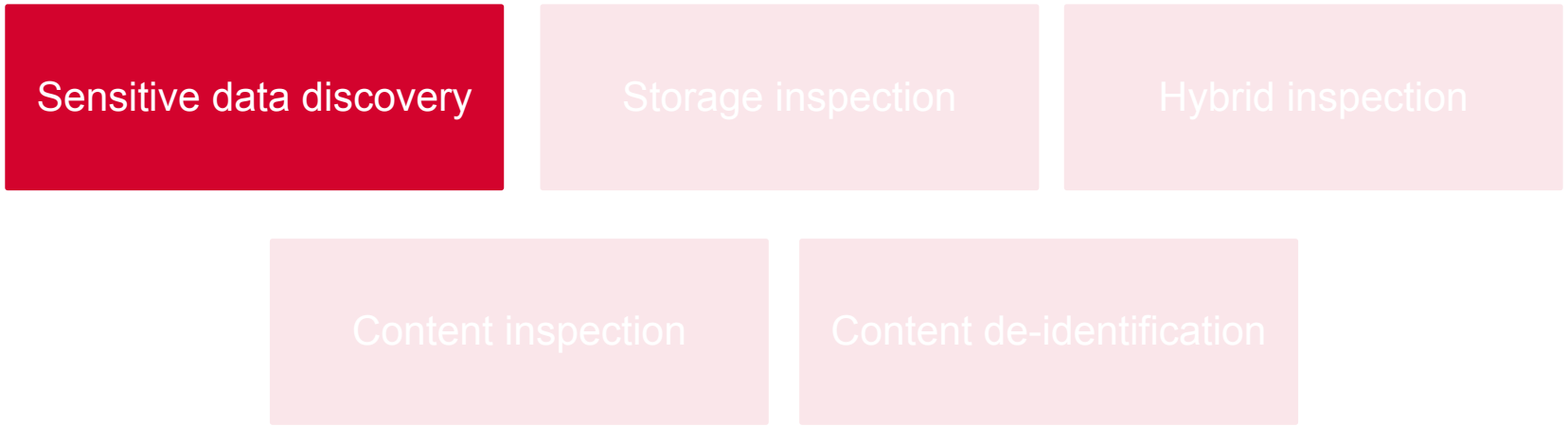
A fully managed service designed to help you discover, classify, and protect your valuable data assets.

The Cloud Data Loss Prevention APIs are now part of this family of managed services. Provides API access to all the services for sensitive data protection.

Sensitive Data Protection

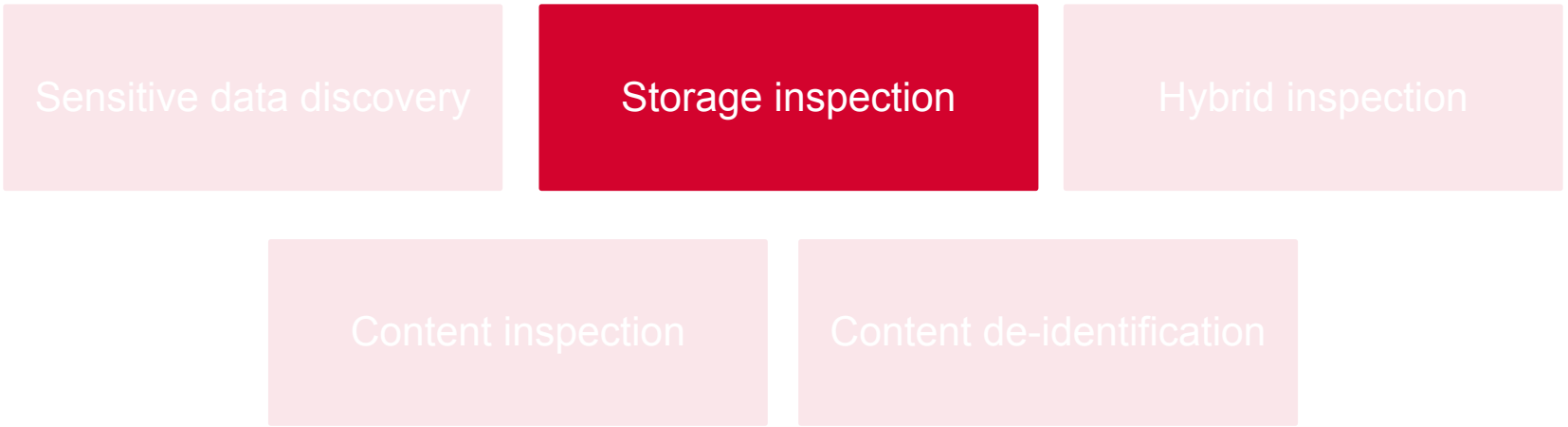


Sensitive Data Protection



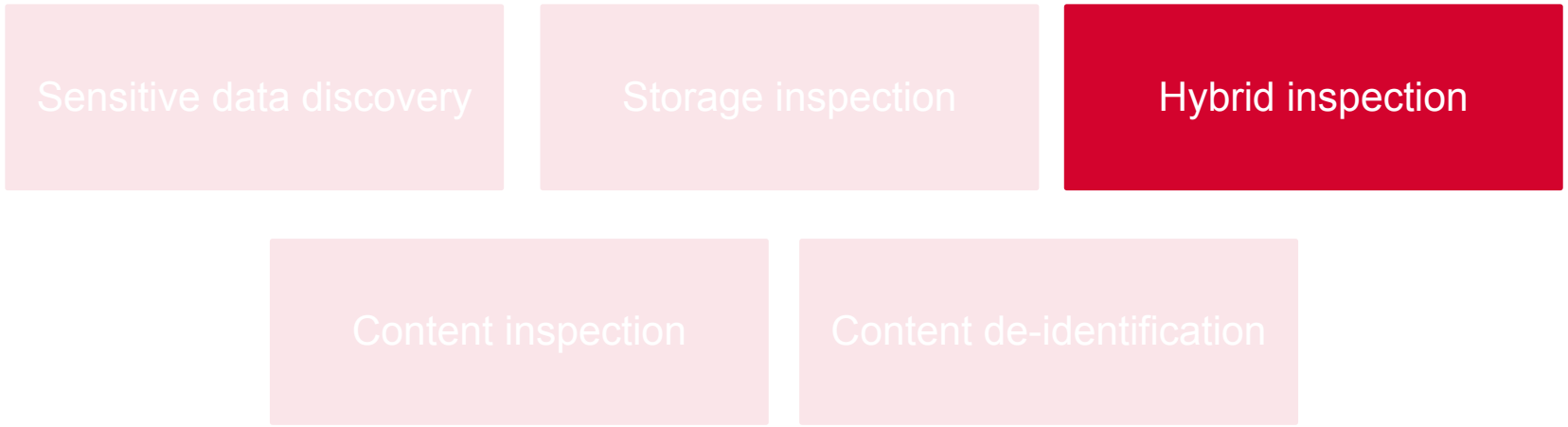
Scan for sensitive data stored in your databases and data warehouses. Use scan configurations to specify what data you are looking for. Constructs data profiles that help you discover sensitive data

Sensitive Data Protection



Scan for and find data **stored in Google Cloud Storage** in unstructured formats e.g chat logs.

Sensitive Data Protection



Scan for and find data stored **outside of Google Cloud** in unstructured formats e.g chat logs.

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Perform data inspection in near real time - used to integrate into custom workloads, applications, or pipelines

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Masking, tokenizing, or de-identifying sensitive data in near real time - used to integrate into custom workloads, applications, or pipelines

Security

A fintech company's fraud detection team needs to analyze a live stream of financial transactions to identify suspicious patterns. The raw transaction events, by an API, contain sensitive data, including full credit card numbers.

For compliance with payment card industry regulations (PCI DSS), the raw credit card numbers must not be stored at all on the Google Cloud

What is the most appropriate and secure Google Cloud solution?

A. Create a streaming Dataflow job that reads from the API. Within the pipeline, use a Cloud Data Loss Prevention (DLP) transformation to inspect the data and tokenize the credit card numbers. Write the transformed, de-identified records to BigQuery.

B. B. Stream the raw data directly into a staging table in BigQuery. Use BigQuery's column-level security and IAM policies to prevent analysts from viewing the column that contains the credit card numbers.

C. Create a streaming Dataflow job that reads from the API and pushes the data to Pub/Sub. Write a Cloud Function that triggers on new Pub/Sub messages. The function should write transaction records containing credit card numbers to a highly restricted, encrypted Cloud Storage bucket, and write all other records to BigQuery.

D. Create a Dataflow job that uses a custom Regular Expression (RegEx) to find and replace number patterns that look like credit card numbers. Load the result into BigQuery.

Security

A fintech company's fraud detection team needs to analyze a live stream of financial transactions to identify suspicious patterns. The raw transaction events, by an API, contain sensitive data, including full credit card numbers.

For compliance with payment card industry regulations (PCI DSS), the raw credit card numbers must not be stored at all on the Google Cloud

What is the most appropriate and secure Google Cloud solution?

A. Create a streaming Dataflow job that reads from the API. Within the pipeline, use a Cloud Data Loss Prevention (DLP) transformation to inspect the data and tokenize the credit card numbers. Write the transformed, de-identified records to BigQuery.

B. B. Stream the raw data directly into a staging table in BigQuery. Use BigQuery's column-level security and IAM policies to prevent analysts from viewing the column that contains the credit card numbers.

C. Create a streaming Dataflow job that reads from the API and pushes the data to Pub/Sub. Write a Cloud Function that triggers on new Pub/Sub messages. The function should write transaction records containing credit card numbers to a highly restricted, encrypted Cloud Storage bucket, and write all other records to BigQuery.

D. Create a Dataflow job that uses a custom Regular Expression (RegEx) to find and replace number patterns that look like credit card numbers. Load the result into BigQuery.

Security

A financial services company is landing sensitive customer transaction data in a Cloud Storage bucket before loading it into BigQuery for analytics. For compliance with industry regulations, they must manage and be able to rotate the encryption keys used for this data at rest.

A critical requirement is that BigQuery must be able to transparently read the data from the Cloud Storage bucket without a separate, manual decryption step in the pipeline.

What is the recommended Google Cloud practice to meet these requirements?

A. Before uploading, use an application to encrypt each file with a key stored in Cloud KMS. The BigQuery load job must then be configured to call the KMS API to decrypt each file.

B. Create a cryptographic key in Cloud KMS. Configure the Cloud Storage bucket's default encryption settings to use this Customer-Managed Encryption Key (CMEK). Grant the BigQuery service account the necessary role on the KMS key.

C. Generate your own AES-256 key on-premises. When uploading files to Cloud Storage, provide this key with the request to use the Customer-Supplied Encryption Key (CSEK) feature.

D. Store the encryption key in Secret Manager. Write a Cloud Function that triggers on new file uploads, reads the key from Secret Manager, and re-writes the file with the encryption applied.

Security

A financial services company is landing sensitive customer transaction data in a Cloud Storage bucket before loading it into BigQuery for analytics. For compliance with industry regulations, they must manage and be able to rotate the encryption keys used for this data at rest.

A critical requirement is that BigQuery must be able to transparently read the data from the Cloud Storage bucket without a separate, manual decryption step in the pipeline.

What is the recommended Google Cloud practice to meet these requirements?

A. Before uploading, use an application to encrypt each file with a key stored in Cloud KMS. The BigQuery load job must then be configured to call the KMS API to decrypt each file.

B. Create a cryptographic key in Cloud KMS. Configure the Cloud Storage bucket's default encryption settings to use this Customer-Managed Encryption Key (CMEK). Grant the BigQuery service account the necessary role on the KMS key.

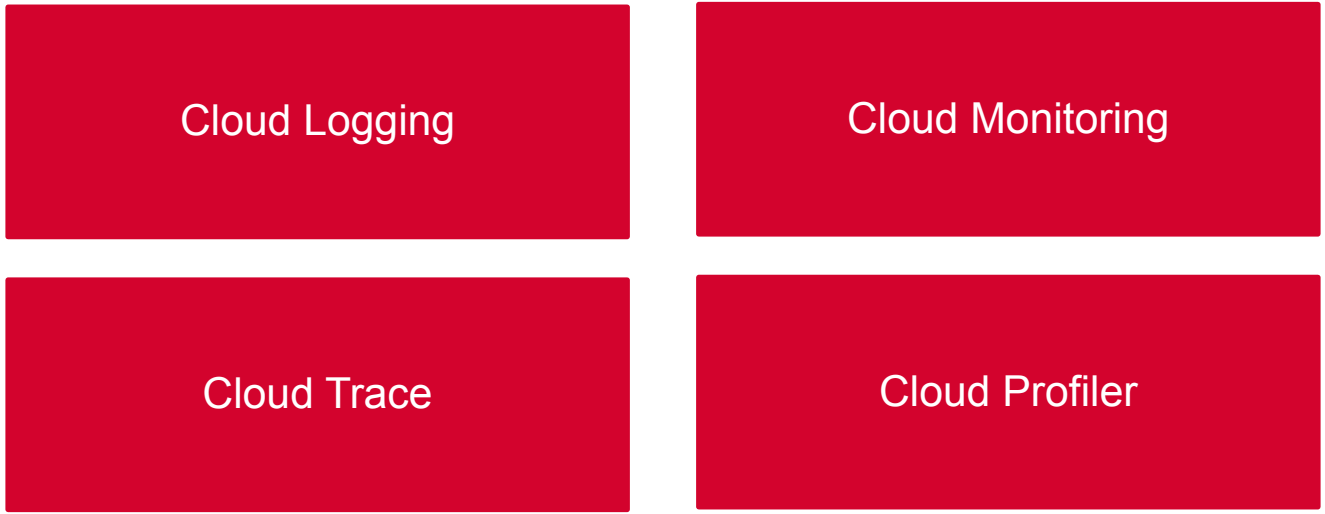
C. Generate your own AES-256 key on-premises. When uploading files to Cloud Storage, provide this key with the request to use the Customer-Supplied Encryption Key (CSEK) feature.

D. Store the encryption key in Secret Manager. Write a Cloud Function that triggers on new file uploads, reads the key from Secret Manager, and re-writes the file with the encryption applied.

O'REILLY®

Observability

Observability on the Google Cloud



Observability on the Google Cloud

Cloud Logging

Cloud Monitoring

Cloud Trace

Cloud Profiler

Cloud Logging

- Centralized service that collects, stores, and analyzes logs from all your apps and infra
  - GCE, Cloud Run, GKE etc.
  - On-prem and other clouds (via Ops Agent)
  - Custom application logs
- Useful for immediate troubleshooting and debugging of your application

Cloud Logging Sinks

- Can automatically route logs to other destinations for long-term storage and analysis
  - Cloud Storage: for compliance and audit
  - BigQuery: for analytics using SQL
  - Pub/Sub: for real-time even-driven workflows

Observability on the Google Cloud

Cloud Logging

Cloud Monitoring

Cloud Trace

Cloud Profiler

Cloud Monitoring

- Central dashboard for performance, uptime, and health across all infra and apps
- View predefined and custom metrics, configure alerts for errors
- Can send uptime checks to applications to verify availability
- Can define SLOs (Service Level Objectives) to measure reliability of systems

Observability on the Google Cloud

Cloud Logging

Cloud Monitoring

Cloud Trace

Cloud Profiler

Cloud Trace

- Used to find latency bottlenecks in applications
  - Answers the question “Why is my application request slow?”
- Visualizes the entire request journey (trace) as it travels through different microservices and function calls
- Useful for debugging micro services - see how they interact and how data flows between them

Observability on the Google Cloud

Cloud Logging

Cloud Monitoring

Cloud Trace

Cloud Profiler

Cloud Profiler

- A continuous, low-overhead profiler that runs in production without slowing down your applications
- Can answer the question “Which lines of my code are consuming the most CPU or memory?”
- Helps you find inefficient code to improve performance and significantly reduce infra costs

Observability

A company runs critical nightly batch processing jobs on a Google Kubernetes Engine (GKE) cluster. The runtime for these jobs has been increasing, and they are now at risk of violating their Service Level Agreement (SLA).

The cluster was initially provisioned without any integrated observability suite, so there is no historical performance data available. You need to collect performance metrics (like CPU and memory utilization) from the running pods to diagnose the bottleneck. The solution must not require re-deploying the application or disrupting the currently running jobs.

What is the most direct and least disruptive way to accomplish this?

A. Execute a `gcloud container clusters update` command on the existing cluster to enable Cloud Operations for GKE. Then, use Cloud Monitoring to analyze the performance metrics of the pods.

B. Cordon and drain the existing GKE nodes one by one. Replace them with new nodes from a new node pool that has Cloud Operations for GKE enabled, allowing the jobs to be rescheduled onto new, monitored nodes.

C. Write a script that repeatedly uses `kubectl top` and `kubectl logs` to manually sample performance data from the pods while the job is running. Store the output in a text file for later analysis.

D. Deploy a third-party monitoring agent, like Prometheus, to each node in the cluster using a Kubernetes DaemonSet. Configure the agent to collect metrics and send them to Cloud Monitoring.

Observability

A company runs critical nightly batch processing jobs on a Google Kubernetes Engine (GKE) cluster. The runtime for these jobs has been increasing, and they are now at risk of violating their Service Level Agreement (SLA).

The cluster was initially provisioned without any integrated observability suite, so there is no historical performance data available. You need to collect performance metrics (like CPU and memory utilization) from the running pods to diagnose the bottleneck. The solution must not require re-deploying the application or disrupting the currently running jobs.

What is the most direct and least disruptive way to accomplish this?

A. Execute a `gcloud container clusters update` command on the existing cluster to enable Cloud Operations for GKE. Then, use Cloud Monitoring to analyze the performance metrics of the pods.

B. Cordon and drain the existing GKE nodes one by one. Replace them with new nodes from a new node pool that has Cloud Operations for GKE enabled, allowing the jobs to be rescheduled onto new, monitored nodes.

C. Write a script that repeatedly uses `kubectl top` and `kubectl logs` to manually sample performance data from the pods while the job is running. Store the output in a text file for later analysis.

D. Deploy a third-party monitoring agent, like Prometheus, to each node in the cluster using a Kubernetes DaemonSet. Configure the agent to collect metrics and send them to Cloud Monitoring.

Observability

Your company's security team needs to meet SOX compliance requirements by retaining all Admin Activity audit logs for a period of five years. These logs track administrative changes to your Google Cloud resources and must be stored in a secure, immutable, and cost-effective manner for potential future audits.

The solution must capture only these specific audit logs from your entire Google Cloud Organization.

What is the most appropriate and direct way to configure this?

A. Enable Data Access audit logs for all services to ensure maximum visibility. Create a sink to export these high-volume logs to a BigQuery table for detailed analysis. Get BigQuery partitions to expire in 5 years

B. Upgrade the default log bucket to use Log Analytics. Run a scheduled query every day to view the Admin Activity logs from the past 24 hours to ensure they are captured. Make sure you can query 5 years worth of data

C. Create a log sink at the organization level. Configure the sink's filter to select only Admin Activity audit logs and set the destination to a Cloud Storage bucket configured with a 5-year retention lock.

D. Create a log-based alert in Cloud Monitoring that triggers whenever an Admin Activity log is generated. Configure the alert to send a notification to a Pub/Sub topic for real-time tracking and storage. Store messages for 5 years

Observability

Your company's security team needs to meet SOX compliance requirements by retaining all Admin Activity audit logs for a period of five years. These logs track administrative changes to your Google Cloud resources and must be stored in a secure, immutable, and cost-effective manner for potential future audits.

The solution must capture only these specific audit logs from your entire Google Cloud Organization.

What is the most appropriate and direct way to configure this?

A. Enable Data Access audit logs for all services to ensure maximum visibility. Create a sink to export these high-volume logs to a BigQuery table for detailed analysis. Get BigQuery partitions to expire in 5 years

B. Upgrade the default log bucket to use Log Analytics. Run a scheduled query every day to view the Admin Activity logs from the past 24 hours to ensure they are captured. Make sure you can query 5 years worth of data

C. Create a log sink at the organization level. Configure the sink's filter to select only Admin Activity audit logs and set the destination to a Cloud Storage bucket configured with a 5-year retention lock.

D. Create a log-based alert in Cloud Monitoring that triggers whenever an Admin Activity log is generated. Configure the alert to send a notification to a Pub/Sub topic for real-time tracking and storage. Store messages for 5 years