Google Cloud

Partner Certification Academy

Professional Cloud Architect

Google Cloud

pls-academy-pca-student-slides-5-2301

The information in this presentation is classified:

Google confidential & proprietary

⚠️ This presentation is shared with you under NDA.

- Do **not** record or take screenshots of this presentation.
- Do **not** share or otherwise distribute the information in this presentation with anyone **inside** or **outside** of your organization.

Thank you!

Google Cloud

Session logistics

- When you have a question, please:
  - Click the Raise hand button in Google Meet.
  - Or add your question to the Q&A section of Google Meet.
  - Please note that answers may be deferred until the end of the session.

- These slides are available in the Student Lecture section of your Qwiklabs classroom.

- The session is **not recorded.**

- Google Meet does not have persistent chat.
  - If you get disconnected, you will lose the chat history.
  - Please copy any important URLs to a local text file as they appear in the chat.

Google Cloud

Professional Cloud Architect (PCA)

Professional Cloud Architects enable organizations to leverage Google Cloud technologies. With a thorough understanding of cloud architecture and Google Cloud, they design, develop, and manage robust, secure, scalable, highly available, and dynamic solutions to drive business objectives.

The Professional Cloud Architect certification exam assesses your ability to:

- ✓ Design and plan a cloud solution architecture
- ✓ Analyze and optimize technical and business processes
- ✓ Manage and provision the cloud solution infrastructure
- ✓ Manage implementations of cloud architecture
- ✓ Design for security and compliance
- ✓ Ensure solution and operations reliability

Google Cloud https://cloud.google.com/certification/cloud-architect

What is a PCA supposed to know/do?

Professional Cloud Architect Certification Page
https://cloud.google.com/certification/cloud-architect

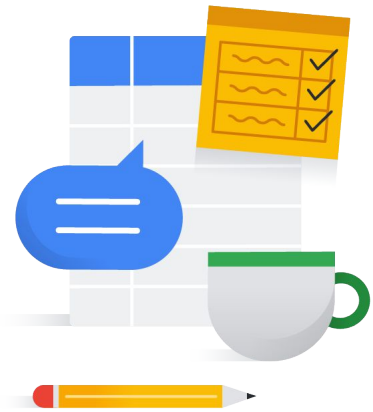
Professional Cloud Architect Exam Guide
https://cloud.google.com/certification/guides/professional-cloud-architect/

Professional Cloud Architect Sample Questions
https://docs.google.com/forms/d/e/1FAIpQLSdvf8Xq6m0kvyIoysdr8WZYCG32WHENStftiHTSdtW4ad2-0w/viewform

Google Cloud Adoption Framework + Whitepaper
https://cloud.google.com/adoption-framework
https://services.google.com/fh/files/misc/google_cloud_adoption_framework_whitepaper.pdf

Program issues or concerns?

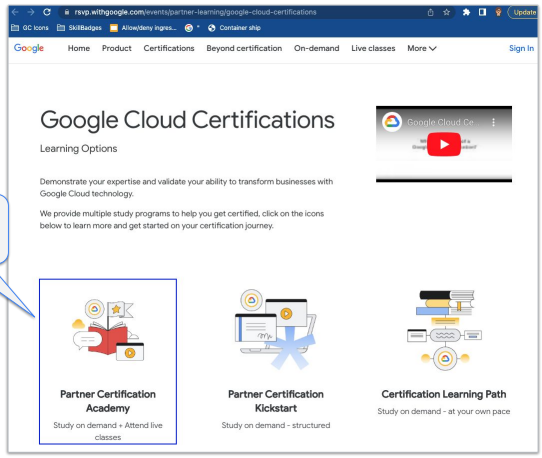
- Problems with **accessing** Cloud Skills Boost for Partners
  - partner-training@google.com
- Problems with **a lab** (locked out, etc.)
  - support@qwiklabs.com
- Problems with accessing Partner Advantage
  - https://support.google.com/googlecloud/topic/9198654



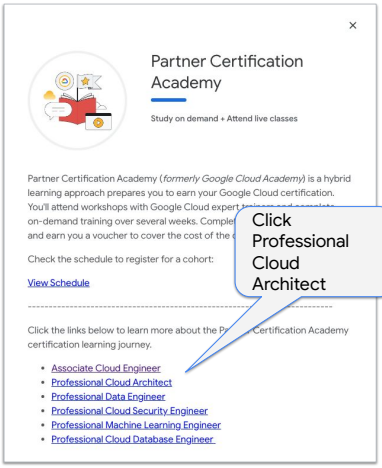
Google Cloud

Learning Path - Partner Certification Academy Website

Go to: https://rsvp.withgoogle.com/events/partner-learning/google-cloud-certifications



Click here



Click Professional Cloud Architect

Google Cloud

PARTNER CERTIFICATION ACADEMY

Professional Cloud Architect

On-demand Course
Hands-on Skill Badge
Hands-on Lab
Resource
Live Virtual Workshop
required to earn exam voucher

Click to register

Pre-Work

- Professional Cloud Architect
- Certification Learning Path: Professional Cloud Architect

Module 1

- A Tour of Google Cloud Hands-on Labs
- Preparing for your Professional Cloud Architect Journey (pre-work)
- Technical Workshop 1
- Technical Workshop 2
- Google Cloud Platform Fundamentals - Core Infrastructure
- Essential Cloud Infrastructure: Foundation
- Google Cloud security foundations guide (white paper)

Module 2

- Technical Workshop 3
- Technical Workshop 4
- Essential Cloud Infrastructure: Core Services
- Service Accounts and Roles: Fundamentals

Needed for Exam Voucher

Google Cloud

PARTNER CERTIFICATION ACADEMY

Professional Cloud Architect

On-demand Course
Hands-on Skill Badge
Hands-on Lab
Resource
Live Virtual Workshop
required to earn exam voucher

Click to register

Module 3

- Technical Workshop 5
- Technical Workshop 6
- Elastic Google Cloud Infrastructure: Scaling and Automation
- HTTP Load Balancer with Cloud Armor
- Create an Internal Load Balancer
- Network Tiers - Optimizing Network Spend
- Perform Foundational Infrastructure Tasks in Google Cloud
- Create and Manage Cloud Resources

Module 4

- Technical Workshop 7
- Technical Workshop 8
- Getting Started with Google Kubernetes Engine
- Setting up a Private Kubernetes Cluster
- Optimize Costs for Google Kubernetes Engine
- Deploy and Manage Cloud Environments with Google Cloud

Module 5

- Technical Workshop 9
- Technical Workshop 10
- Reliable Google Cloud Infrastructure: Design and Process
- Getting Started with Cloud KMS
- User Authentication: Identity-Aware Proxy
- Securing Virtual Machines using Beyond Corp Enterprise
- Set Up and Configure a Cloud Environment in Google Cloud
- Automating Infrastructure on Google Cloud with Terraform

Module 6

- Logging, Monitoring, and Observability in Google Cloud
- Technical Workshop 11
- Exam Readiness Workshop
- Cloud Architecture: Design, Implement, and Manage
- Reporting Application Metrics into Cloud Monitoring
- Monitoring Multiple Projects with Cloud Monitoring

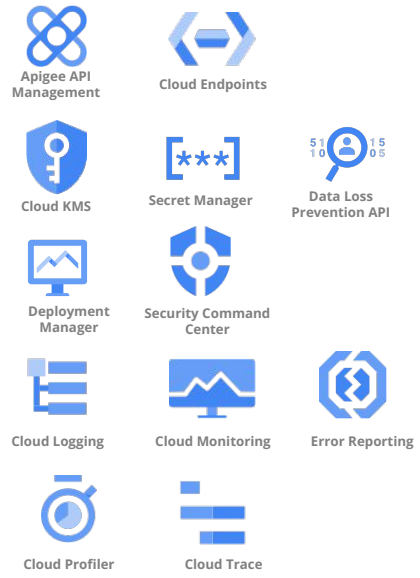
Needed for Exam Voucher

02

Key concepts in the on-demand content

Topics in this module

- API Management
- Data security
- Secret Manager
- Data Loss Prevention (DLP)
- VPC Service Controls
- Private Google Access
- Identity Aware Proxy (IAP)
- Microservices
- Infrastructure as Code
- Compliance
- Security Command Center
- Designing for Reliability
- Designing for Disaster Recovery
- Cloud Operations



Google Cloud

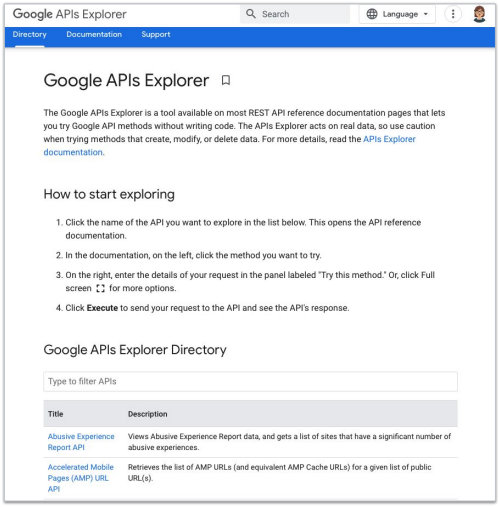
API management...

Google Cloud

APIs (Application programming Interfaces)

- Behind the scenes, all interactions with Google Cloud are done via API calls
  - Google has an APIs Explorer site which allows developers to call an API and see the results
    - Useful in app development
- Organizations often create their own APIs to allow developers to interact with internal systems to create
  - Customer-facing websites
  - Internal applications for inventory management, etc.
- APIs must be secured, documented, versioned, tested, etc.

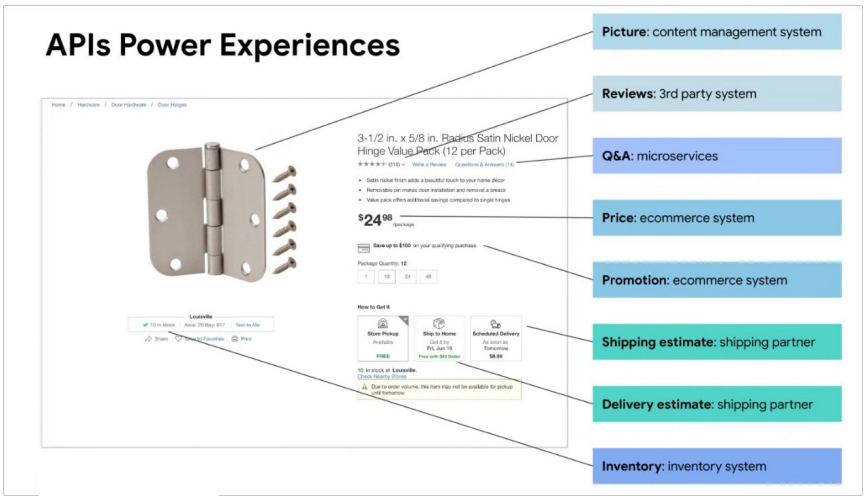
Google Cloud



https://developers.google.com/apis-explorer

APIs - customer use case

- Screenshot from [The Home Depot's website]
- This webpage makes at least 6 different API calls
- If they decide to redesign this [page] and move items around, the APIs don't change



Google Cloud

Proprietary + Confidential

Google Cloud provides two tools, Cloud Endpoints and Apigee, for managing APIs

Both provide tools for:

- User authentication
- Monitoring
- Securing APIs
- Etc.

Both support OpenAPI and gRPC

Google Cloud

Cloud Endpoints

Google Cloud Only
2 million/mth free

Apigee API Platform

Hybrid Deployments
Monetization feature

Apigee
https://docs.apigee.com/

Apigee Youtube:
https://www.youtube.com/watch?v=58smxQu3P5k

Apigee Demo videos
https://cloud.google.com/apigee/demo-ty

Cloud Endpoints
https://cloud.google.com/endpoints

Both Apigee and Cloud Endpoints are built with a goal to manage all your APIs.

Endpoints is an API management gateway that helps you develop, deploy, and manage APIs on any Google Cloud backend. It runs on Google Cloud and leverages a lot of Google's underlying infrastructure.

Apigee is an API management platform built for enterprises, with deployment options on cloud, on-premises, or hybrid. The feature set includes an API gateway, customizable portal for onboarding partners and developers, monetization, and deep analytics around APIs. You can use Apigee for any http/https backends, no matter where they are running (on-premises, any public cloud, etc.).

API Management Tools

- Cloud Endpoints
  - https://cloud.google.com/endpoints
  - Cloud Endpoints Portal
    - Demo: https://endpointsportal.endpoints-portal-demo.cloud.goog
- Apigee API Management
  - https://cloud.google.com/apigee

Google Cloud

From the exam guide

3.1 Designing for security. Considerations include:

- Identity and access management (IAM)
- Resource hierarchy (organizations, folders, projects)
- **Data security (key management, encryption, secret management)**
- Separation of duties (SoD)
- Security controls (e.g., auditing, VPC Service Controls, context aware access, organization policy)
- **Managing customer-managed encryption keys with Cloud Key Management Service**
- Remote access

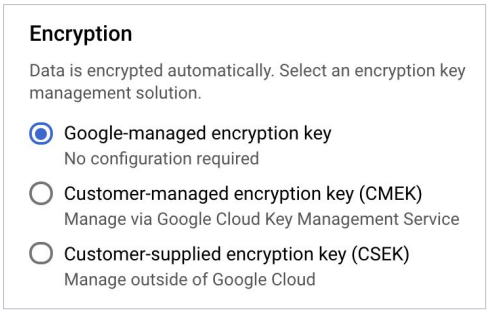
Google Cloud

Data security (key management, encryption, secret management)

Data is **always encrypted** in Google Cloud, at no additional charge.

- Applies to **all storage**, e.g., object storage and block storage (VHD)

Encryption Options



Encryption at rest in Google Cloud

Google Cloud

Encryption at rest in Google Cloud
https://cloud.google.com/docs/security/encryption/default-encryption

Encryption in Transit in Google Cloud
https://cloud.google.com/docs/security/encryption-in-transit

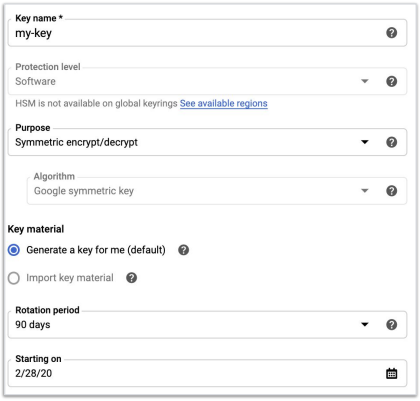
Customer-managed encryption keys (CMEK)
https://cloud.google.com/kms/docs/cmek

Customer-Supplied Encryption Keys
https://cloud.google.com/docs/security/encryption/customer-supplied-encryption-keys

Encryption options

- Default - Google Cloud manages encryption
- Customer-managed encryption keys (CMEK)
  - Encryption occurs after Cloud Storage receives data but before the data is written to disk
    - Keys are managed through Cloud Key Management Service (KMS)
- Customer-supplied encryption keys (CSEK)
  - Keys created and managed externally to Google Cloud
    - Keys act as an additional encryption layer on top of the standard encryption

- Client-side encryption (external to Google Cloud)
  - Data is sent to Cloud Storage already encrypted
    - Data also undergoes server-side encryption by Google



Cloud KMS

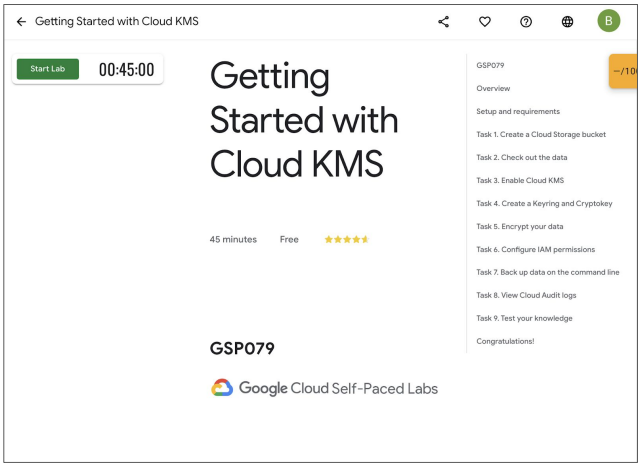
Google Cloud

Related products

Cloud HSM is a cloud-hosted Hardware Security Module (HSM) service that allows you to host encryption keys and perform cryptographic operations in a cluster of FIPS 140-2 Level 3 certified HSMs. https://cloud.google.com/kms/docs/hsm

Support for external keys with EKM:
Encrypt data in BigQuery and Compute Engine with encryption keys that are stored and managed in a third-party key management system that's deployed outside Google's infrastructure. External Key Manager allows you to maintain separation between your data at rest and your encryption keys while still leveraging the power of cloud for compute and analytics.
https://cloud.google.com/security-key-management

Suggested Lab



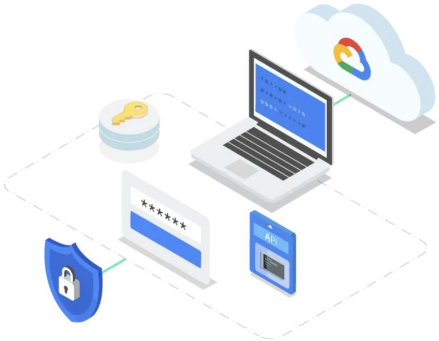
https://partner.cloudskillsboost.google/catalog_lab/368

Google Cloud

Data security (key management, encryption, secret management)

Secret Manager

- Secure storage system for API keys, passwords, certificates, and other sensitive data
- Single source of truth to manage, access, and audit secrets across Google Cloud
- Every interaction generates an audit log
- Simple lifecycle management with versioning and the ability to pin requests to the latest version of a secret
- IAM used to control access



Google Cloud

Secret Manager

Secret Manager
https://cloud.google.com/secret-manager

Managing Cloud Run secrets securely
https://www.youtube.com/watch?v=JIE89dneaGo

Data security - other topics

Cloud DLP (Data Loss Protection)

Fully managed service to discover, classify, and protect sensitive data

[Improving security, compliance, and governance]

Before

ID	Job Title	Phone	Comments
359740	Senior Engineer	307-964-0673	Please email them at jane@imadethisup.com
981587	VP, Engineer	713-910-6787	none
394091	Lawyer	692-398-4146	Updated phone to: 692-398-4146
986941	Senior Ops Manager	294-967-5508	none
490456	Junior Ops Manager	791-954-3281	Tried to verify account with their SSN 222-44-5555

After

ID (FPE)	Job Title	Phone	Comments
438422	Engineer	307-###-####	Please email them at [Found Email]
530375	Engineer	713-###-####	none
496534	Lawyer	692-###-####	Updated phone to: 692-###-####
242348	Ops	294-###-####	none
593887	Ops	791-###-####	Tried to verify account with their SSN [Found SSN]

Google Cloud

Cloud Data Loss Prevention
https://cloud.google.com/dlp

Improving security, compliance, and governance with cloud-based DLP data discovery
https://cloud.google.com/blog/products/identity-security/modern-dlp-for-cloud-data-discovery

From the exam guide

3.1 Designing for security. Considerations include:

- Identity and access management (IAM)
- Resource hierarchy (organizations, folders, projects)
- Data security (key management, encryption, secret management)
- Separation of duties (SoD)
- Security controls (e.g., auditing, VPC Service Controls, context aware access, organization policy)
- Managing customer-managed encryption keys with Cloud Key Management Service
- Remote access

Google Cloud

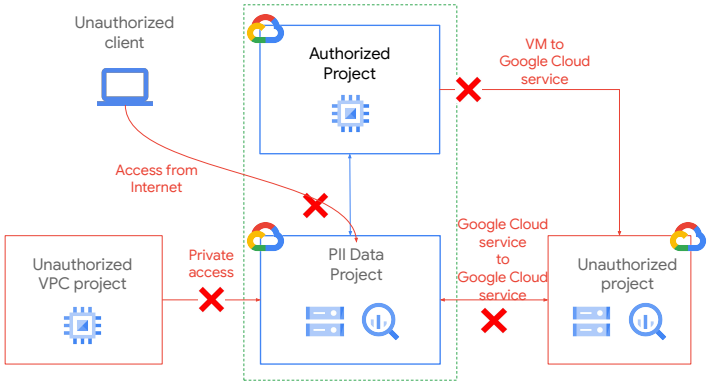
What if you need to prevent VMs or other types of users from accessing managed services in certain projects?

Google Cloud

Security controls (e.g., auditing, VPC Service Controls, context aware access)

Improves your ability to mitigate the risk of data exfiltration from services such as Cloud Storage and BigQuery

- Uses perimeters to define what resources and data need to be protected
- Functions like a firewall for GC APIs
- Choose which projects and services to protect
- Can allow access to the perimeter **based on context,** for example:
  - Access from on-premises networks during working hours
  - Access from specific geo-locations



Google Cloud

- https://cloud.google.com/vpc-service-controls/docs/overview

- **Service perimeters**
  - Defined by **adding projects and services** to include within the perimeter.
  - A **project** can belong to **one perimeter only.**
- In this example, we have a **service perimeter** around **two projects**, including a **VPC network, Cloud Storage and BigQuery.**
- This means that
  - **VMs in the VPC network** can only reach **services within the perimeter.**
  - Accordingly, the **services** within the perimeter **can only communicate with each other**.
- Let's see which **risks** this helps protecting against
  - **Stolen identity / misconfigured IAM**
  - **Data exfiltration through VM**:
    - A malice actor exploiting a **vulnerability in a VM** within the perimeter, attempts to **copy data from the VM** to their own external bucket.
    - The same malice actor can try to copy data **using Cloud Storage or BigQuery built-in commands.**
    - Both attempts would be **blocked by VPC SC**, but **wouldn't be blocked by a VPC FW**.
  - Access to BigQuery / Cloud Storage from **VM outside of the perimeter**

- Additionally, you could **allow access** to the perimeter **based on context,** for example:
  - Access from on-premises networks during working hours
  - Access from specific geo-locations

Security controls - other topics
Private Google Access

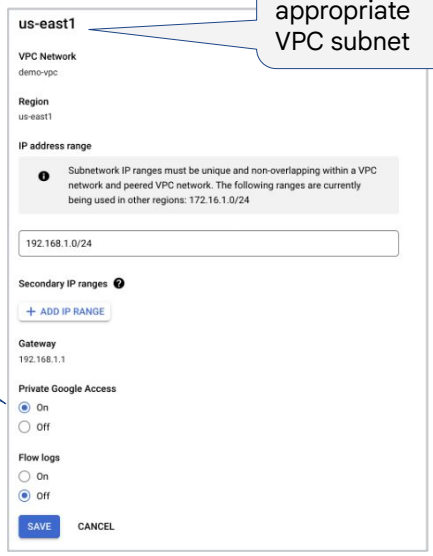
- Certain Google API based managed services do not use VPC Networks
  - For example
    - Cloud Functions
    - Cloud Storage
    - Cloud Run
    - Container/Artifact Registry
    - AI Platform - Video Intelligence, Translation, etc.
- They use **public** service endpoints that specifies the network address of an API service
- For VMs to reach these, they need an external IP
  - External IPs on VMs are not best practice
- Solution
  - Enable Private Google Access in the subnetwork to which the VM is attached

Google Cloud https://cloud.google.com/vpc/docs/private-google-access

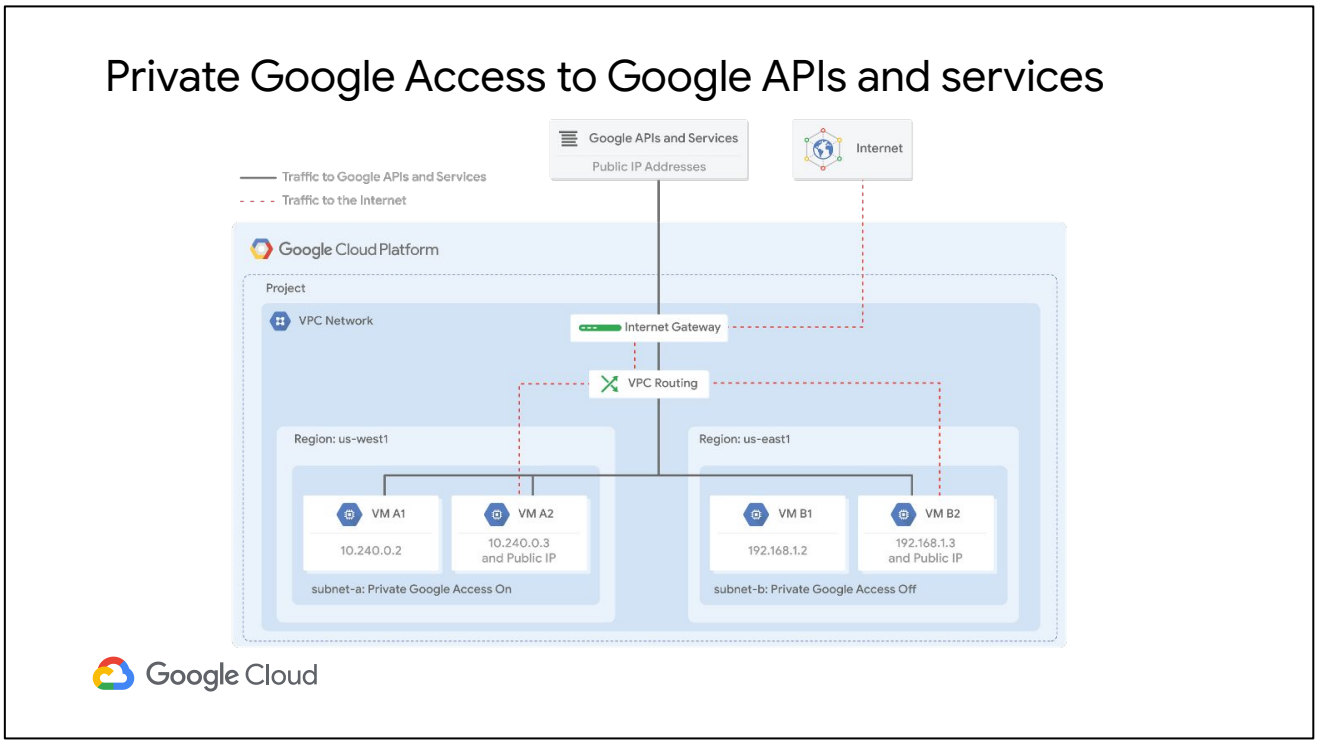
Enabling Private Google Access

Select the appropriate VPC subnet

Enable it



Google Cloud

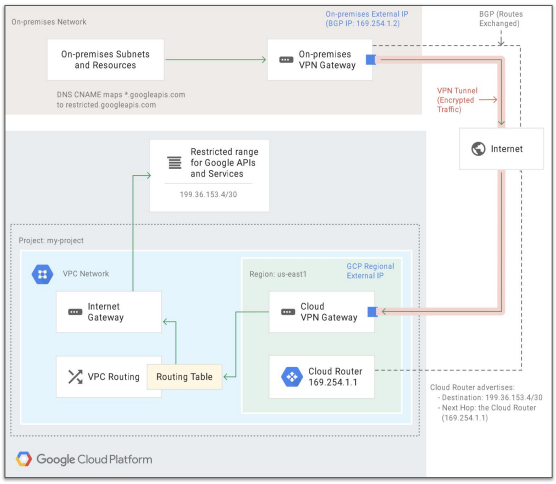


You enable Private Google Access on a subnet-by-subnet basis. As you can see in this diagram, subnet-a has Private Google Access enabled, and subnet-b has it disabled. This allows VM A1 to access Google APIs and services, even though it has no external IP address.

Private Google Access has no effect on instances that have external IP addresses. That's why VMs A2 and B2 can access Google APIs and services. The only VM that can't access those APIs and services is VM B1. This VM has no public IP address, and it is in a subnet where Google Private Access is disabled.

Private Google Access On-Prem

- Allows on-premises hosts to reach Google APIs and services using internal IPs
  - Must use Cloud VPN or Cloud Interconnect
- To enable Private Google Access for on-premises hosts,
  - Configure DNS, firewall rules, and routes in on-premises and VPC networks.
  - **No need to enable Private Google Access for any subnets in your VPC network**



Google Cloud

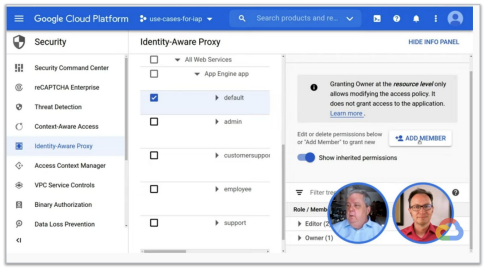
Private Google Access for on-premises hosts
https://cloud.google.com/vpc/docs/private-google-access-hybrid

Security controls - other topics: Identity Aware Proxy (IAP)

- **Service to authenticate users using the Google Identity Service**
  - Requests are restricted to authorized users
  - Plays a key role in Google's BeyondCorp zero trust security model
    - Shift access controls from the network perimeter to individual user attributes (location, device, etc.)
    - Users can work from virtually any location without the need for a traditional VPN
  - **Developers do not have to make any changes to their code**
- Two options
  - IAP for HTTPS - Google Cloud and on-premise
  - IAP for TCP tunneling (later discussion)

Google Cloud

Recommended Youtube video



https://www.youtube.com/watch?v=xM9-FSU5MoY

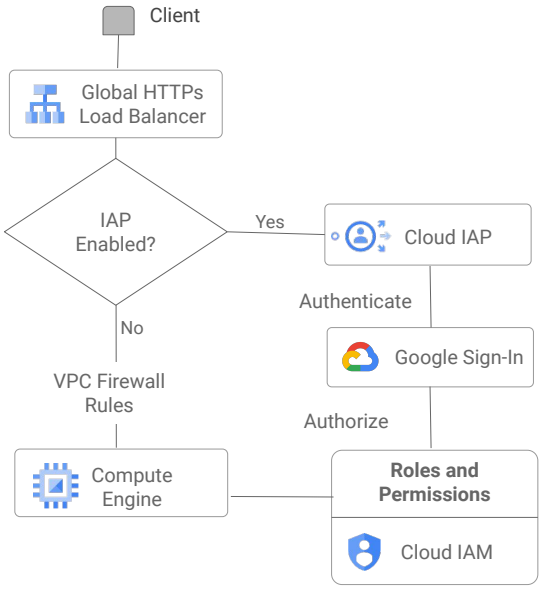
Centralize access to your organization's websites with Identity Aware Proxy (IAP)
Youtube link on slide: https://www.youtube.com/watch?v=xM9-FSU5MoY

Getting started with BeyondCorp: A deeper look into IAP
Youtube: https://www.youtube.com/watch?v=goAWiQGRefw&t=341s

BeyondCorp Enterprise
https://cloud.google.com/beyondcorp-enterprise

Identity Aware Proxy (IAP) for HTTPS

- Integrated with HTTP(s) Load Balancer
  - Users are prompted to sign-in
- Single point of control for managing user access based on
  - Policy configurations, e.g, which group does a user belong to, which URL is accessed, which device is used. etc.
  - User IAM roles
- No VPN needed
- Little to no change to applications
  - Don't need to implement own authentication for each application
- Available for Compute Engine, Kubernetes Engine, App Engine and Cloud Run



Google Cloud

Identity-Aware Proxy
https://cloud.google.com/iap

IAP overview:
https://cloud.google.com/iap/docs/concepts-overview

IAP Enabling external identities:
https://cloud.google.com/iap/docs/enable-external-identities

Cloud IAP enables context-aware access to VMs via SSH and RDP without bastion hosts
https://cloud.google.com/blog/products/identity-security/cloud-iap-enables-context-aware-access-to-vms-via-ssh-and-rdp-without-bastion-hosts/

Youtube: Centralize access to your organization's websites with Identity Aware Proxy (IAP)
https://www.youtube.com/watch?v=xM9-FSU5MoY

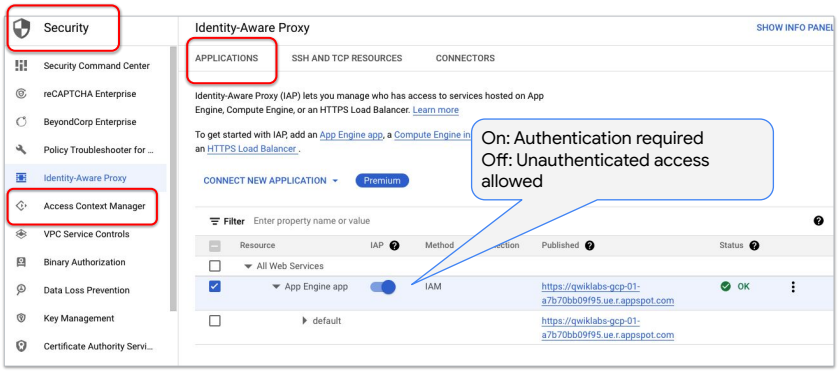
Key points:

- Cloud IAP provides a **much simpler administration process** and with **reduced operational overhead** than more **traditional VPN** solutions.

    - There is **no VPN to implement** or VPN clients to install and maintain.
    - It also makes the end **user experience more streamlined** as the user no longer has to launch the VPN client and sign into the VPN.
- **How it works**
    - Enables an **app-level access control model** instead of a **network-level control model**
    - Access only possible through **IAP** by users with the **right IAM role**
    - **Authentication** with **Google**, including **2FA**
    - **Authorization** with **Cloud IAM** - users need *IAP-secured Web App User* on the resource project
    - Simple **integration** with **HTTPS Load Balancer** for **Compute Engine/GKE**
        - **Little** to no app/code **change required**
- **Best practice:** apply **VPC firewalls** to **ensure traffic** goes through the **LB**
    - **IAP** only **protects traffic** going **through the LB**
    - **App-serving ports** are still **accessible**

Enabling IAP for HTTP(S) Applications - Task 1

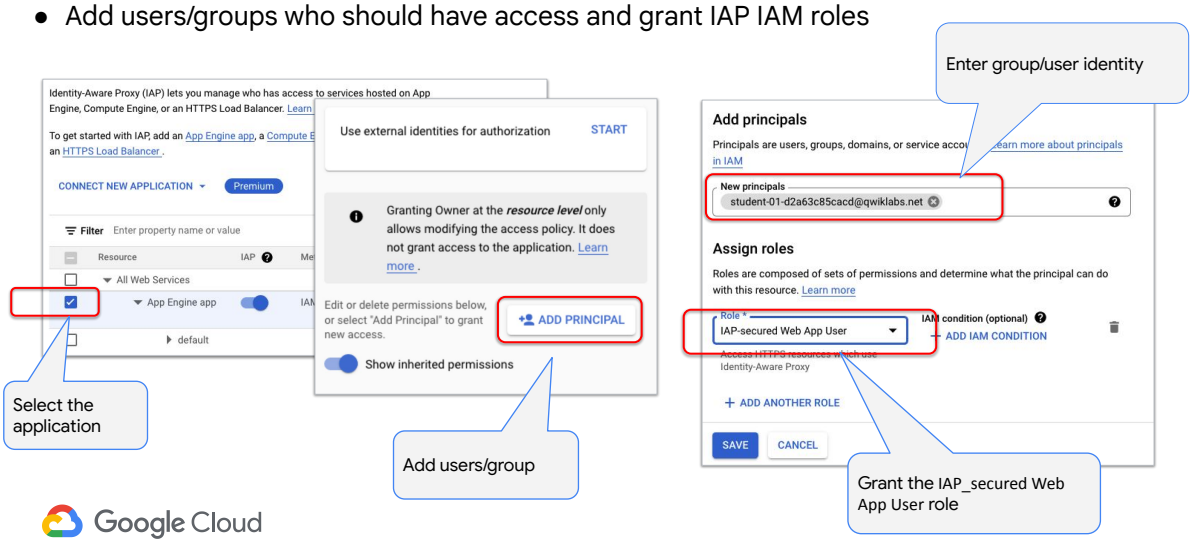
- Create the website
- Enable IAP API and provide information about your privacy policy, etc. to be displayed on login screen
- Select the web app and turn on/off IAP



Google Cloud

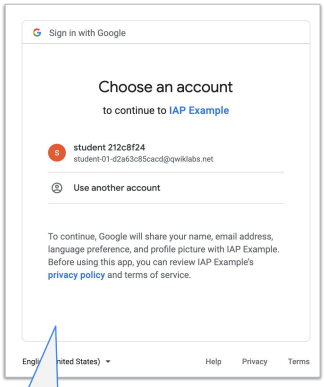
Enabling IAP for HTTP(S) Applications - Task 2

- Add users/groups who should have access and grant IAP IAM roles

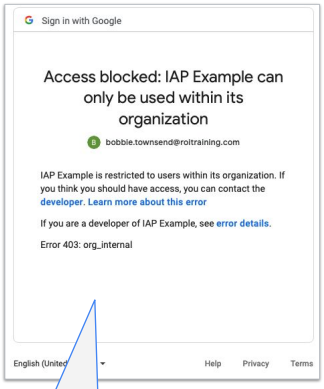


Google Cloud

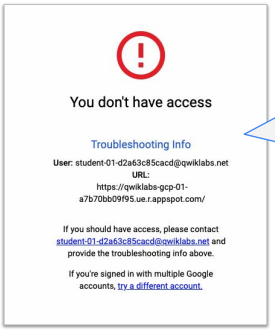
IAP for HTTP(S) - Login examples



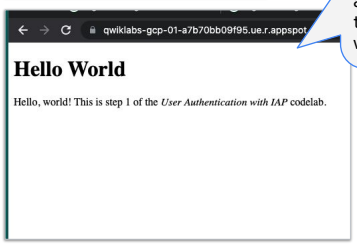
Login Screen



External person blocked



Internal person without "IAP_secured Web App User" role blocked

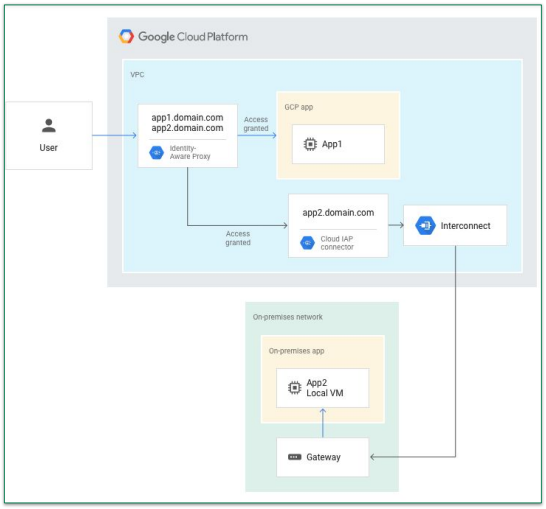


Internal person with "IAP_secured Web App User" role allowed through to website

Google Cloud

Identity Aware Proxy (IAP) for on-premise apps

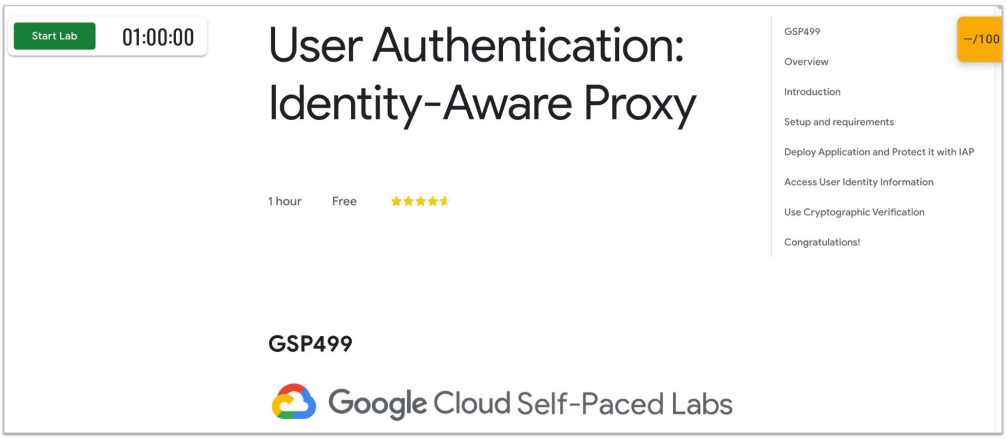
- Manage access to HTTP(S)-based apps outside of Google Cloud
- Requests are authenticated and authorized
  - If destined to an endpoint on-prem, an **IAP connector** forwards the request through a site-to-site connection using Cloud Interconnect from Google Cloud to the on-premises network.



Google Cloud

https://cloud.google.com/iap/docs/cloud-iap-for-on-prem-apps-overview

Suggested lab: User Authentication: Identity-Aware Proxy

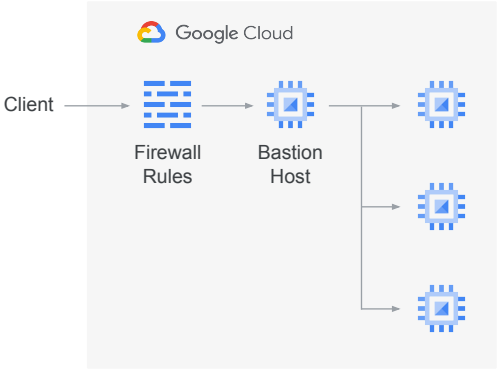


https://partner.cloudskillsboost.google/catalog_lab/1741

Google Cloud

Techniques to access VMs without external IPs

- **Use a bastion host**
  - A temporary VM with a public IP
  - Connect to it, then connect to the private IPs of instances from there
    - **Con: Still have a VM with a public IP**
- **Use Cloud VPN/Interconnect**
  - Connect from on-prem network to the Google Cloud network via private IP address
    - **Con: Developers are limited as to where they must be (i.e, connected to the network/VPN)**
- Use **Identity-Aware Proxy for TCP forwarding** to forward a SSH/RDP connection to a remote instance without the need for a VPN connection



Google Cloud

There are several options available for securely communicating with VMs that do not have public IP addresses. These services do not have a public IP address normally because they are deployed to be consumed by other instances in the project or maybe through **Dedicated Interconnect** options. However, for those instances that do not have an external IP address, it can be a requirement to gain external access, for example for updates or patches to be applied. The options for accessing the VMs include:

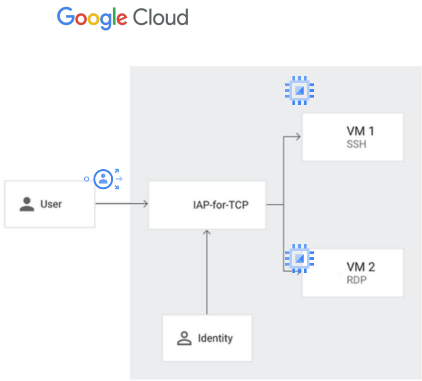
Bastion Host: A bastion host provides an external facing point of entry into a network containing private network instances. This host provides a single point of secure access and can be stopped to disable inbound SSH. This allows the connection to VMs without having to configure firewall rules. Typical hardening initial steps for a bastion host include limiting the CIDR range of source IP addresses that can communicate with the host and configuring firewall rules to only allow SSH to private VM addresses from the bastion host.

Identity-Aware Proxy: IAP was mentioned earlier in this section. Using SSH with IAP's TCP forwarding feature wraps the SSH connection inside HTTPS. The TCP forwarding then sends it to the remote instance.

From the options above, the preferred approach for inbound SSH is to use IAP, which is a built-in service, rather than configuring a one-off bastion host.

Identity-Aware Proxy for TCP Forwarding

- Tunnel TCP traffic to instances without exposing them to the public internet
  - Traffic between client and IAP is wrapped in HTTPS
- Can be used instead of a bastion host by using Cloud SDK
  - SSH
    - *gcloud compute ssh user@instance --zone <zone>*
  - RDP
    - Start the tunnel on a local port with a *gcloud* command and use Windows Remote Desktop to connect to the port



Google Cloud

Cloud IAP enables context-aware access to VMs via SSH and RDP without bastion hosts
https://cloud.google.com/blog/products/identity-security/cloud-iap-enables-context-aware-access-to-vms-via-ssh-and-rdp-without-bastion-hosts/

Identity-Aware Proxy for TCP Forwarding:
https://cloud.google.com/iap/docs/tcp-forwarding-overview

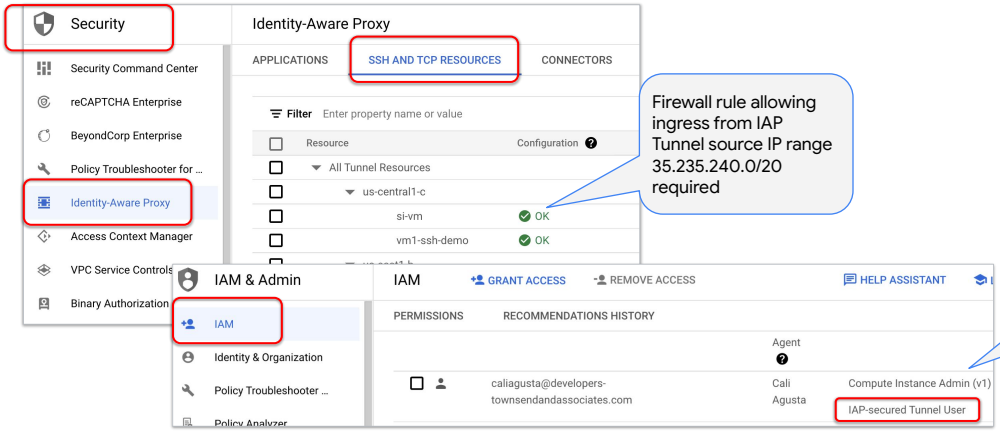
Using IAP for TCP forwarding
https://cloud.google.com/iap/docs/using-tcp-forwarding

Identity-Aware Proxy for TCP service can be used to tunnel TCP traffic to Compute Engine VMs that do not have external ip addresses
Access is authenticated with Google and authorized with IAM

A very common use case is connecting to Compute Engine VMs using SSH or RDP and can serve as a Bastion host replacement

- Simply use the Cloud SDK to connect to a VM
- The command is for ssh is
  gcloud compute ssh your-user-name and the zone in which the VM resides

Identity-aware proxy for TCP Tunneling - Example



Firewall rule allowing ingress from IAP Tunnel source IP range 35.235.240.0/20 required

Group/User must have **IAP-secured Tunnel User** role

```
$ gcloud compute ssh caliagusta@vm-iap-tunnel-demo --zone=us-central1-a
External IP address was not found; defaulting to using IAP tunneling.
//output removed for brevity
Last login: Thu Sep 29 19:36:20 2022 from 35.235.244.34
caliagusta@vm-iap-tunnel-demo:~$
```

Using shell to connect to the VM from a remote computer

Google Cloud

Suggested lab: Securing Virtual Machines using BeyondCorp Enterprise (BCE)

- Topics include
  - IAP TCP forwarding
  - Linus/Windows instances
  - Firewall rules
  - IAP Desktop

https://partner.cloudskillsboost.google/catalog_lab/4947

Google Cloud

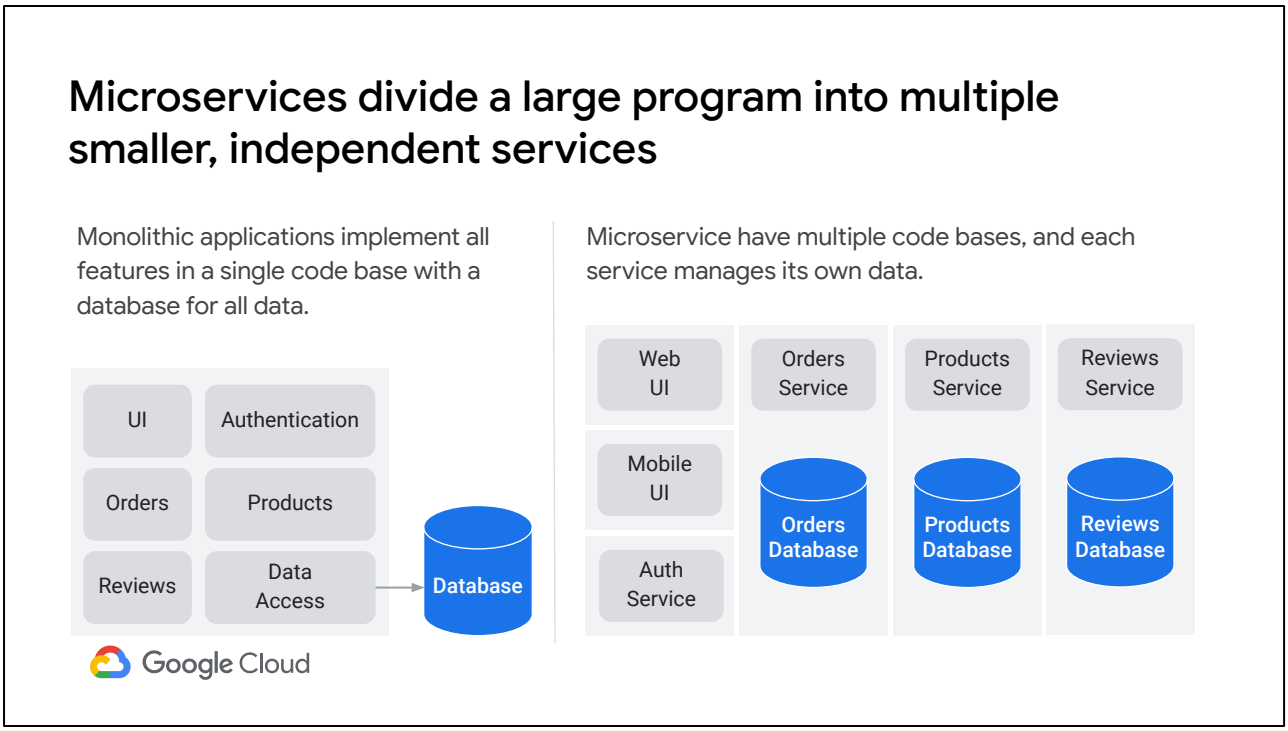
Summary of security controls discussed in this section

- **Private google access**
  - Allows VM without public IPs to access managed services
- **VPC Service Controls**
  - Creates a perimeter around which projects and which services can be accessed from within the perimeter
    - No other access is allowed
- **Identity Aware Proxy (IAP)**
  - IAP for HTTP(S) applications
    - Authentication for apps accessed with HTTPS load balancer
  - IAP - TCP forwarding
    - Authentication for GCE instances accessed via SSH or RDP
  - IAP - On Premise
    - Authentication for apps accessed with Google HTTPS load balancer, then routed to on-premise servers

Google Cloud

Microservices...

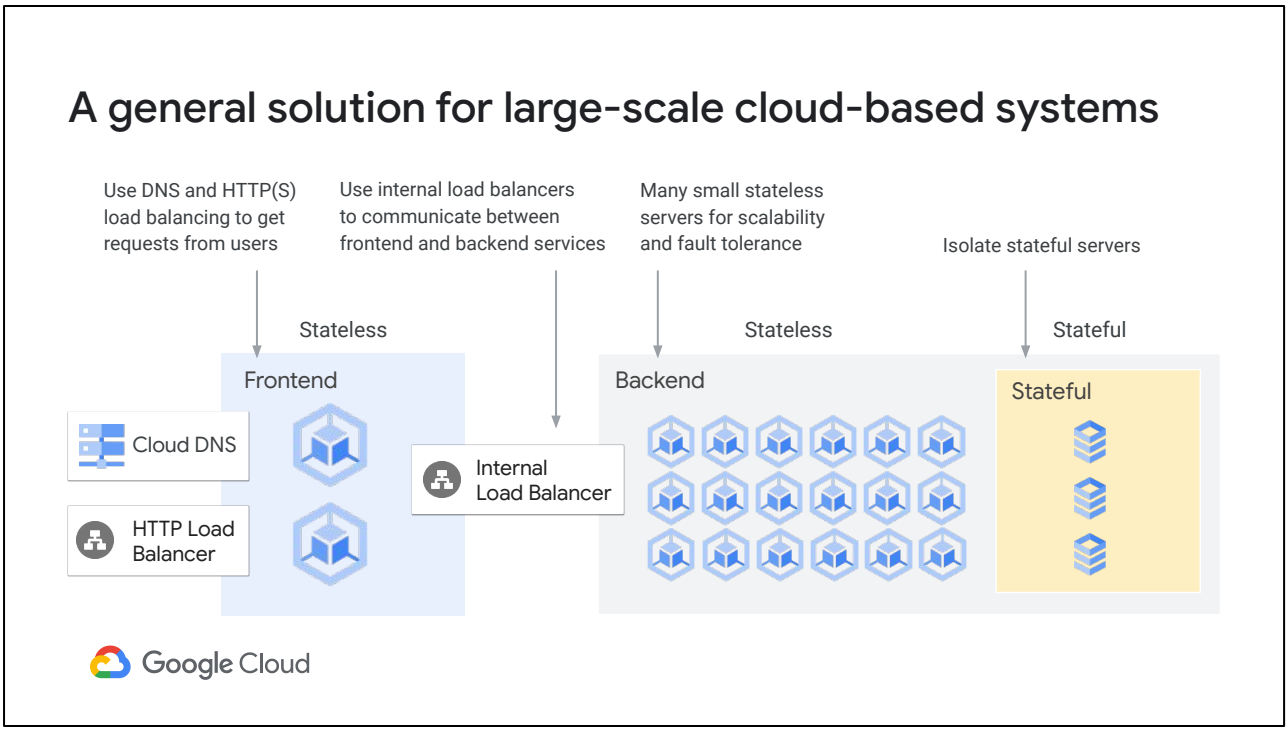
Google Cloud



Microservice architecture is the current trend. It is important to ensure that there is a good reason to select this architecture. The primary reason is to enable teams to work independently and deliver through to production at their own cadence. This supports scaling the organization: adding more teams increases speed. There is also the additional benefit of being able to scale the microservices independently based on their requirements.

Architecturally, a monolith or microservices should be modular components with clearly defined boundaries. With a monolith, the deployment is the grouping of the components, whereas with microservices, the individual components are deployable. Google Cloud provides several services for deploying microservices from App Engine, Cloud Run, GKE, and Cloud Functions. Each offers different levels of granularity and control and will be discussed later in the course.

To achieve independence on services, each service should have its own datastore. This lets the best datastore solution for that service be selected. Google Cloud is strong here, with a wide selection of datastores.



The slide displays a general solution that shows the separation of the frontend and backend processing stages. A load balancer distributes the load between the backend and frontend services. This allows the backend to scale if it needs to keep up with the demand from the frontend. In addition, the stateful servers/services are also isolated.

The layout above allows a large part of the application to make use of the scalability and fault tolerance of Google Cloud services as stateless services. By isolation of the stateful servers and services, the challenges of scaling and upgrading are limited to a subset of the overall set of services.

Customer case study - How switching to microservices and GKE helped Kitabisa meet fundraising demands

- Kitabisa's mission is to raise funds when disasters like earthquakes strike
  - During times of crisis, the crush of donations pushed internal systems to a breaking point
- During their transition to Google Cloud
  - Their monolith fundraising application was converted to microservices running on GKE and now scales as needed
  - Their monolithic database was migrated to Cloud SQL with multiple read replicas
- Read more about this transformation here

Google Cloud

How Kitabisa re-structured its fundraising platform to drive "kindness at scale" on Google Cloud

The result surpassed our expectations. During the following year's Ramadan season, we gained a 50% boost in computing resources to easily handle escalating crowdfunding demands on our system.

How Kitabisa re-structured its fundraising platform to drive "kindness at scale" on Google Cloud
https://cloud.google.com/blog/products/application-development/how-kitabisa-uses-google-cloud-gke-cloud-sql-and-proxysql/

More about Kitabisa
https://kitabisa.com/

Infrastructure as Code (IaC)...

Google Cloud

Moving to the cloud requires a mindset change

On-Premises

- Buy machines.
- Keep machines running for years.
- Prefer fewer big machines.
- Machines are capital expenditures.

Cloud

- Rent machines.
- Turn machines off as soon as possible.
- Prefer lots of small machines.
- Machines are monthly expenses.

Google Cloud

The on demand, pay-per-use model of cloud computing is a different model to traditional on-premises infrastructure provisioning. Resources can be allocated to best meet demand in a timely manner, and the cloud supports experimentation and innovation by providing immediate access to an ever-increasing range of services.

Treat infrastructure as disposable in the cloud

- Don't fix broken machines.
- Don't install patches.
- Don't upgrade machines.
- If you need to fix a machine, delete it and re-create a new one.

- Many tools exist for automating infrastructure creation
  - Terraform
  - Deployment Manager on Google Cloud
  - CloudFormation on AWS
  - Resource Manager on Azure

Google Cloud

The key term is infrastructure as code (IaC). The provisioning, configuration, and deployment activities should all be automated.

Having the process automated minimizes risks, eliminates manual mistakes, and supports repeatable deployments and scale and speed. Deploying one or one hundred machines is the same effort.

Costs can be reduced by provisioning ephemeral environments, such as test environments that replicate the production environment.

Infrastructure as code (IaC) allows quick provisioning and removing of infrastructures

- Build an infrastructure when needed.
- Destroy the infrastructure when not in use.
- Create identical infrastructures for dev, test, and prod
  - Use templates to create different types of resources per group
- Can be part of a CI/CD pipeline.
- Templates are the building blocks for disaster recovery procedures.

Google Cloud

In essence, infrastructure as code allows for the quick provisioning and removing of infrastructures.

The on-demand provisioning of a deployment is extremely powerful. This can be integrated into a continuous integration pipeline that smoothes the path to continuous deployment.

Automated infrastructure provisioning means that the infrastructure can be provisioned on demand, and the deployment complexity is managed in code. This provides the flexibility to change infrastructure as requirements change. And all the changes are in one place. Infrastructure for environments such as development and test can now easily replicate production and can be deleted immediately when not in use. All because of infrastructure as code.

Several tools can be used for IaC. Google Cloud supports Terraform, where deployments are described in a file known as a configuration. This details all the resources that should be provisioned. Configurations can be modularized using templates, which allows the abstraction of resources into reusable components across deployments.

In addition to Terraform, Google Cloud also provides support for other IaC tools, including:

- Deployment Manager
- Chef

- Puppet
- Ansible
- Packer

Deployment Manager YAML Templates

```
resources:
# Configure a VM
- name: devops-vm
  type: compute.v1.instance
  properties:
    zone: us-central1-a
    machineType: zones/us-central1-a/machineTypes/f1-micro
    disks:
    - deviceName: boot
      type: PERSISTENT
      boot: true
      autoDelete: true
      initializeParams:
        sourceImage:
projects/debian-cloud/global/images/family/debian-8
    # Add VM to default network and give it an external IP
    networkInterfaces:
    - network: global/networks/default
      accessConfigs:
        - name: External NAT
          type: ONE_TO_ONE_NAT
```

1. Create a virtual machine
2. VM properties
3. Create a boot disk from an image
4. Need a network

Google Cloud

Deployment Manager is Google Cloud's IaC tool
https://cloud.google.com/deployment-manager/docs

Shown here is a simple Deployment Manager template written in YAML. The root element at the top is resources.

Resources is a collection. Notice the minus sign before the name element. In YAML, the minus sign denotes one item in a collection. In this case, we are creating a VM named devops-vm.

The VM has a collection of properties, one of which is a collection of disks. Lastly, at the bottom we are assigning the machine to a network.

Managing deployments with `gcloud`

- Once the template is built, use `gcloud` to deploy it
  - Simple creation of infrastructure
  - Easy changes with update command
  - Delete will remove everything created in the reverse order
- Examples:
  - `gcloud deployment-manager deployments create devops-deployment --config deployment-manager-config.yaml`
  - `gcloud deployment-manager deployments list`
  - `cloud deployment-manager deployments update example-deployment --config deployment-manager-config.yaml --preview`
  - `gcloud deployment-manager deployments delete devops-deployment`

Google Cloud

The gcloud commands used with Deployment Manager are pretty straightforward. Like all gcloud commands you specify the service, collection verb, name, and parameters.

So all the commands begin with gcloud deployment-manager deployments. The verbs are create, list, update, and delete. Lastly, specify the deployment name and the parameters.

When running an update, only resources that have changed in the template will be updated. When running a delete command, it is smart enough to delete resources in the reverse order in which they were created. Thus, you don't get an error when deleting a resource that is used by another resource. As an example, an Instance Group cannot be deleted if a Load Balancer back end is using it. Deployment Manager handles those types of issues for you.

Terraform susports multi-cloud environments

- Pre-installed in Cloud Shell
- Supports both a native syntax named HCL plus JSON-compatible syntax as shown here

```
1  resource "google_compute_instance" "default" {
     name         = "flask-vm"
     machine_type = "f1-micro"
     zone         = "us-west1-a"
2    tags         = ["ssh"]

     boot_disk {
       initialize_params {
         image = "debian-cloud/debian-9"
       }
     }

3    # Install Flask
     metadata_startup_script = "sudo apt-get update; sudo
   apt-get install -yq build-essential python-pip rsync; pip
   install flask"

4    network_interface {
       network = "default"

       access_config {
         // Ephemeral public IP
       }}}
```

1. Create a virtual machine
2. Create a boot disk from an image
3. Run a script at startup
4. Network interface to default VPC & external IP

Google Cloud

Terraform
https://www.terraform.io/

Managing Terraform deployments

- Once the template is built, use terraform commands to make changes
  - `terraform init` : get plugins if needed
  - `terraform plan` : checks for errors in the terraform file
  - `terraform apply` : applies the file and creates the resources
  - `terraform delete` : deletes the resources

Google Cloud

Cloud Foundation Toolkit

- Ready made IaC templates which reflect best practices, in both
  - Deployment Manager
  - Terraform
- Can be used off-the-shelf to quickly build a repeatable enterprise-ready foundation in Google Cloud
  - Can easily update the foundation as needs change

Google Cloud

Some of the example templates

Resource	Description
Autoscaler	Create a Compute Engine autoscaler.
Backend Service	Create a global or regional backend service.
Bastion Host	Create a bastion host, which you can use to access other servers in the same network.
BigQuery	Create a BigQuery dataset and table.
Cloud Functions	Create a Cloud Functions function.
Cloud Router	Create a Cloud Router.
Cloud Spanner	Create a Spanner instance and database.
Cloud SQL	Create a Cloud SQL instance with databases and users.
Cloud Tasks	Create a Cloud Tasks task and task queue.
Dataproc	Create a Dataproc cluster.
Cloud DNS managed zone	Create a managed zone in Cloud DNS.
Cloud DNS records	Create Cloud DNS records using recordsets.
External Load Balancer	Create an HTTP(S), SSL Proxy, or TCP Proxy external load balancer.
Firewall rules	Create firewall rules for your network.

Cloud Foundation Toolkit

Cloud Foundation Toolkit
https://cloud.google.com/foundation-toolkit

From the exam guide

3.2 Designing for compliance. Considerations include:

A. Legislation (e.g., health record privacy, children's privacy, data privacy, and ownership)
B. Commercial (e.g., sensitive data such as credit card information handling, personally identifiable information [PII])
C. Industry certifications (e.g., SOC 2)
D. Audits (including logs)

Google Cloud

Compliance Resource Center

- Google Cloud conforms to global and industry-specific compliance standards
  - GDPR in the EU
  - FedRAMP and SOX in the US
  - HIPPA for the healthcare industry in US
  - PCI for banking
  - Many others...
- **As a Professional Cloud Architect, you should be aware of compliance standards that are required for applications you are building**
  - See: https://cloud.google.com/security/compliance/

Data security
Data residency
Data privacy

Google Cloud

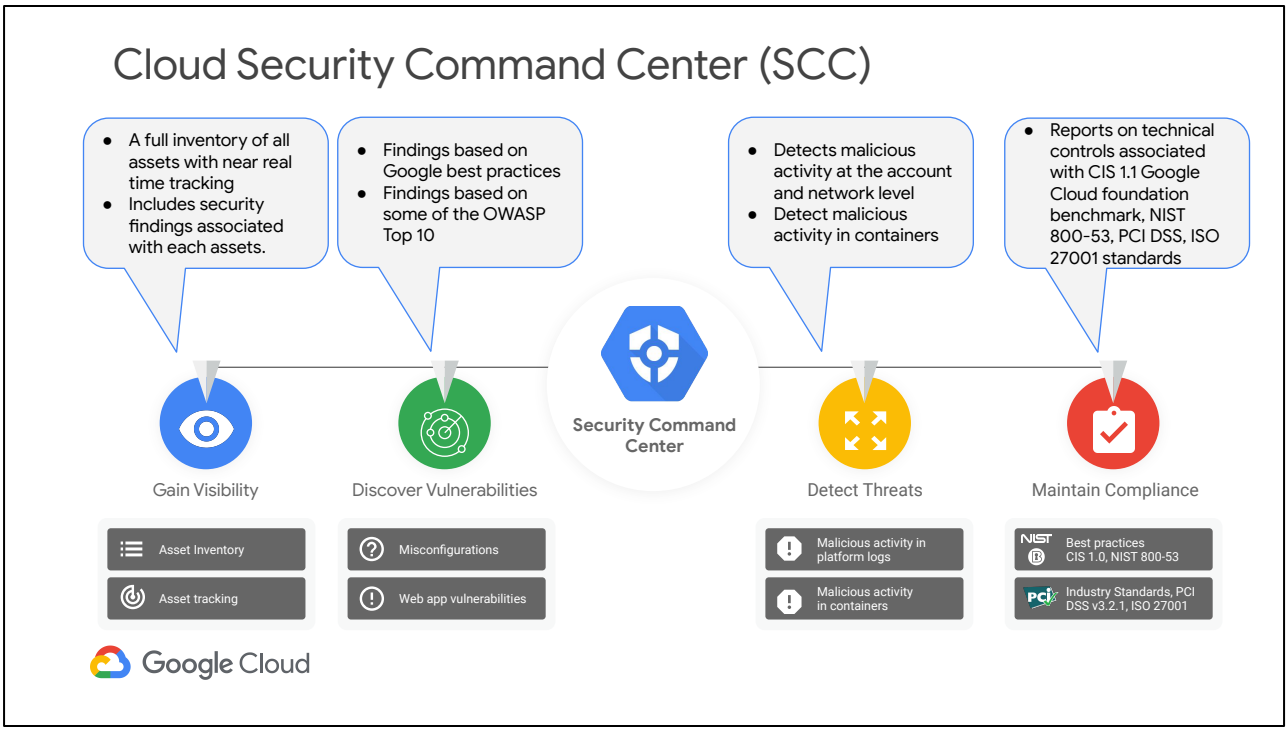
Google Cloud Security, Privacy and Compliance
https://cloud.google.com/security

Compliance Resource Center
https://cloud.google.com/security/compliance/

Google Cloud Compliance Examples

- Shared responsibility
  - PCI (Payment Card Industry) compliance
    - https://cloud.google.com/solutions/pci-dss-compliance-in-gcp
  - HIPAA compliance (US)
    - https://cloud.google.com/security/compliance/hipaa-compliance
  - GDPR compliance
    - https://cloud.google.com/security/gdpr

Google Cloud



Security Command Center
https://cloud.google.com/security-command-center

Security Command Center conceptual overview:
https://cloud.google.com/security-command-center/docs/concepts-security-command-center-overview

OWASP Top 10
https://www.hacksplaining.com/owasp

NIST 800-53: Definition and Tips for Compliance:
https://www.varonis.com/blog/nist-800-53

Benefits

Gain centralized visibility and control
Understand the number of projects you have, what resources are deployed, and manage which service accounts have been added or removed.

Fix misconfigurations and compliance violations
Identify security misconfigurations and compliance violations in your Google Cloud assets and resolve them by following actionable recommendations.

Detect threats targeting your Google Cloud assets

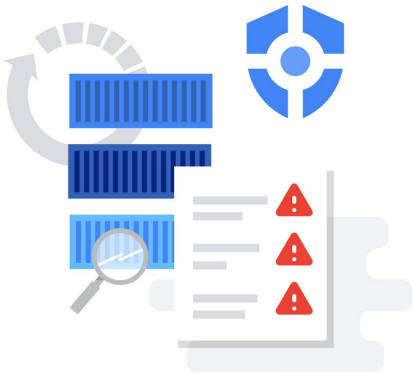
Uncover threats targeting your resources using logs and powered by Google's unique threat intelligence; use kernel-level instrumentation to identify potential compromises of containers.

Recommended Videos

Six Part Video Series - Security Command Center (June 2020 - Dec 2020)
Security Talks: Improve your security posture with the Security Command Center (Nov 18, 2020)

Container Threat Detection

- Detects the most common container runtime attacks and alerts you to any suspicious activity
- Key findings:
  - Suspicious Binary Executions
    - Discover if a binary that was not a part of the original container image is executed
  - Suspicious Library Loaded
    - Detect when a library that was not part of the original container image is loaded
  - Reverse Shell
    - Allow attackers to communicate from a compromised workload to an attacker controlled machine and perform malicious activities, e.g., create botnets



Google Cloud

One of the services you can leverage is Container Threat Detection. It is a built-in service in Security Command Center Premium tier. It detects the most common container runtime attacks and alerts you to any suspicious activity.

Here are the key findings that are identified by Container Threat Detection:

Suspicious Binary Executions: Container Threat Detection can see when a binary that was not part of the original container image is executed, and triggers a finding, indicating that an attacker may have control of the workload and that they are executing suspicious software such as malware or cryptocurrency mining software.

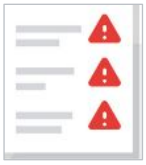
Suspicious Library Loaded: Container Threat Detection can also detect when a library that was not part of the original container image is loaded—a possible sign that the attacker has control of the workload and that they are executing arbitrary code.

Reverse Shell: Container Threat Detection monitors for processes that get started with stream redirection to a remote connected socket. An attacker can use a reverse shell to communicate from a compromised workload to an attacker controlled machine and perform malicious activities, for example as part of a botnet.

Container Threat Detection (Dec 1, 2020)
https://cloud.google.com/blog/products/identity-security/container-threat-detection-is-ga

Other Key Features of Security Command Center

- Asset discovery and inventory
  - Discover and view your assets in near-real time
  - Review historical discovery scans to identify new, modified, or deleted assets
  - Understand the number of projects you have, what resources are deployed, and manage which service accounts have been added or removed.
- Fix misconfigurations and compliance violations
  - Continuously evaluate your security posture against multiple compliance standards including
    - Payment Card Industry Data Security Standard
    - Open Web Application Security Project (OWASP) Top Ten
    - National Institute of Standards and Technology 800-53



Google Cloud

Security Command Center Pricing Tiers

Standard:

- Security Health Analytics
  - Detects the highest severity vulnerabilities and misconfigurations
- Web Security Scanner custom scans
  - Custom scans for a subset of OWASP Top Ten security risks for web applications
- Plus more - see docs

Premium (Standard features plus):

- Everything Standard offers
- Event Threat Detection
- Container Threat Detection
- Security Health Analytics (w/industry best practices and compliance monitoring)
- Web Security Scanner (managed scans)
- Continuous Exports (to Pub/Sub)

Google Cloud

Tier Information
https://cloud.google.com/security-command-center/docs/concepts-security-command-center-overview

Recommended Videos
Six Part Video Series - Security Command Center (June 2020 - Dec 2020)
https://www.youtube.com/watch?v=Q-fzz2P6omQ&list=PLIivdWyY5sqKd-Cu1HZ7v5RiYE8gVsM7P&index=2

Security Talks: Improve your security posture with the Security Command Center (Nov 18, 2020)
https://www.youtube.com/watch?v=4nelxX_1erk&t=20s

Designing for reliability...

Google Cloud

When designing for reliability, consider these key performance metrics

Availability

The percent of time a system is running and able to process requests

- Achieved with fault tolerance.
- Create backup systems.
- Use health checks.

Durability

The odds of losing data because of a hardware or system failure

- Achieved by replicating data in multiple zones.
- Do regular backups.
- Practice restoring from backups.

Scalability

The ability of a system to continue to work as user load and data grow

- Monitor usage.
- Use capacity auto- scaling to add and remove servers in response to changes in load.

Google Cloud

Google Cloud infrastructure reliability guide
https://cloud.google.com/architecture/infra-reliability-guide

Everything you need to know about architecting reliable infrastructure for Google Cloud workloads
https://cloud.google.com/blog/products/infrastructure-modernization/design-reliable-infrastructure-for-workloads-in-google-cloud

Google Cloud Architecture Framework: Reliability
https://cloud.google.com/architecture/framework/reliability

Design for scale and high availability
https://cloud.google.com/architecture/framework/reliability/design-scale-high-availability

Product reliability guides
https://cloud.google.com/architecture/framework/reliability/product-guides

Achieving reliable insights into how the software is functioning is vital. These insights are typically provided by monitoring, logging, and alerting. The performance metrics availability, durability, and scalability all contribute to the reliability of the system.

Availability
To achieve high availability, monitoring is vital. Health checks can detect when an

application reports that it is ok. More detailed monitoring of services using the four golden signals from Google's Site Reliability Engineering book will help determine more insights into the health of a service. Removing single points of failure is also vital for improving availability.

Durability
Ensuring that data is preserved and available is a mixture of replication and backup. Regular restores from backup should be performed to confirm that the process works as expected.

Scalability
Monitoring and autoscaling should be used to respond to variations in load. The metrics for scaling can be the standard CPU/memory, or you can use custom metrics.

When designing for reliability, consider these key performance metrics

Availability

The percent of time a system is running and able to process requests

- Achieved with fault tolerance.
- Create backup systems.
- Use health checks.

Durability

The odds of losing data because of a hardware or system failure

- Achieved by replicating data in multiple zones.
- Do regular backups.
- Practice restoring from backups.

Scalability

The ability of a system to continue to work as user load and data grow

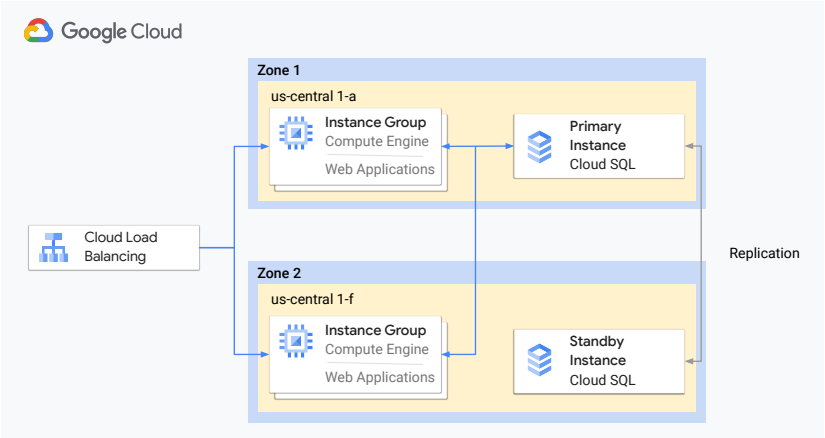
- Monitor usage.
- Use capacity auto- scaling to add and remove servers in response to changes in load.

Google Cloud

Next discussion

When designing for high availability, deploy to multiple zones in a region

- Deploy multiple servers.
- Orchestrate servers with a regional managed instance group.
- Create a failover database in another zone or use a distributed database like Firestore or Spanner.



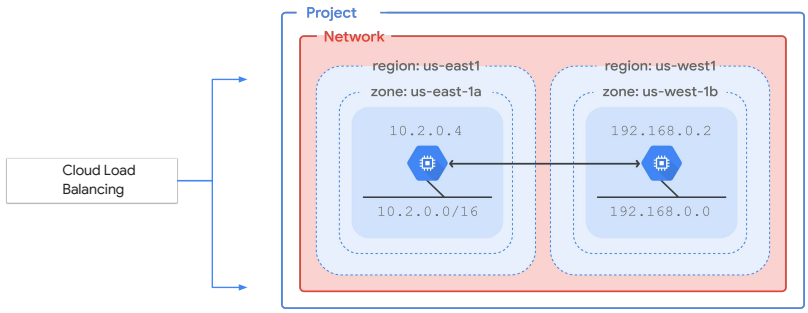
Google Cloud

Design for high availability
https://cloud.google.com/architecture/scalable-and-resilient-apps#physically_distribute_resources

When you're using Compute Engine, for higher availability you can use a regional instance group, which provides built-in functionality to keep instances running. Use auto-healing with an application health check and load balancing to distribute load. For data, the storage solution selected will affect what is needed to achieve high availability. For Cloud SQL, the database can be configured for HA, which provides data redundancy and a standby instance of the database server in another zone.

When designing for globalization, deploy to multiple regions

- Deploy services close to end-users for lowest latency
- Also provides redundancy
  - If one region become unavailable, traffic is automatically routed to the next closest with capacity



Google Cloud

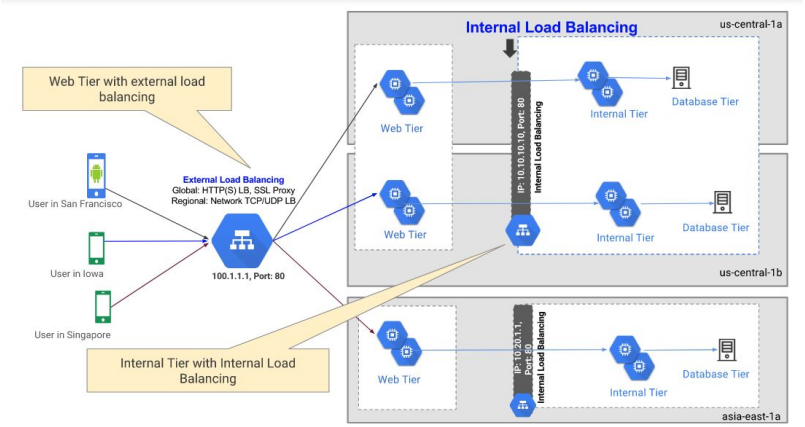
Next, let's look at globalization.

In the previous design we placed resources in different zones in a single region, which provides isolation from many types of infrastructure, hardware, and software failures. Putting resources in different regions as shown on this slide provides an even higher degree of failure independence. This allows you to design robust systems with resources spread across different failure domains.

When using a global load balancer, like the HTTP load balancer, you can route traffic to the region that is closest to the user. This can result in better latency for users and lower network traffic costs for your project.

Load balance at each tier

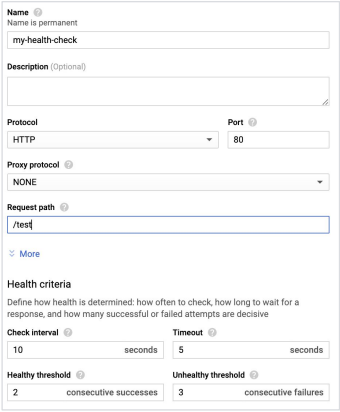
- Load balancing distributes traffic among groups of resources
- This help ensure that individual resources don't become overloaded while others sit idle
- Most load balancers also provide health-checking features to help ensure that traffic isn't routed to unhealthy or unavailable resources



Google Cloud

Create a health check when creating instance groups to enable auto healing

- Create a test endpoint in your service.
  - Used to verify that the service is up, and also that it can communicate with dependent backend database and services.
- If health check fails, the instance group will create a new server and delete the broken one.
- Load balancers also use health checks to ensure that they send requests only to healthy instances.



Google Cloud

When designing for reliability, consider these key performance metrics

Availability

The percent of time a system is running and able to process requests

- Achieved with fault tolerance.
- Create backup systems.
- Use health checks.

Durability

The odds of losing data because of a hardware or system failure

- Achieved by replicating data in multiple zones.
- Do regular backups.
- Practice restoring from backups.

Scalability

The ability of a system to continue to work as user load and data grow

- Monitor usage.
- Use capacity auto- scaling to add and remove servers in response to changes in load.

Next discussion

Google Cloud

Scalability
Monitoring and autoscaling should be used to respond to variations in load. The metrics for scaling can be the standard CPU/memory, or you can use custom metrics.

Configure autoscaling

- Configure autoscaling behavior based on key app metrics, on cost profile, and on defined minimum required level of resources
- Many Google Cloud compute products offer autoscaling
  - Serverless managed services like Cloud Run, Cloud Functions, and App Engine are designed to scale quickly and automatically
  - MIGs can scale based on various inputs, including Cloud Monitoring custom metrics and load-balancer serving capacity
    - Can set minimum and maximum limits on the scaling behavior
  - GKE can autoscale to add or remove nodes based on workload compute requirements
  - Spanner/Bigtable can increase/decrease compute capacity w/autoscaling

Google Cloud

Patterns for scalable and resilient apps
https://cloud.google.com/architecture/scalable-and-resilient-apps

Autoscaling Cloud Spanner:
https://cloud.google.com/architecture/autoscaling-cloud-spanner

Bigtable scaling:
https://cloud.google.com/bigtable/docs/scaling

Memorystore scaling:
https://cloud.google.com/memorystore/docs/redis/scaling-instances#

Storage buckets offer high availability

- Regional buckets are redundant across zones
- Multi/dual regions buckets are redundant across regions/zones

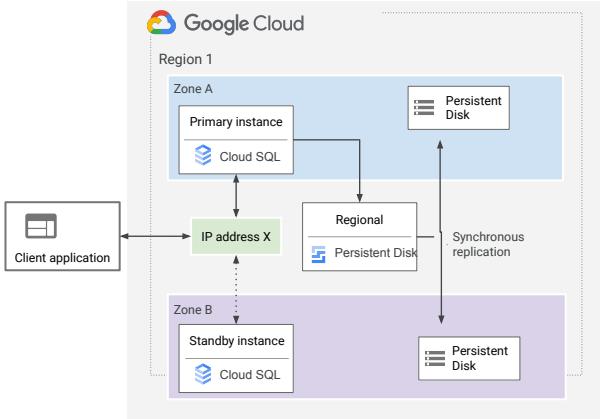
Bucket type	Availability
Multi-region	99.95%
Single region	99.90%
Dual-region	99.95%

Google Cloud

The multi-region availability benefit is a factor of 2—0.1% to 0.05% unavailability. But in terms of time it is not so much. 99.9% availability means 43 minutes 12 seconds per month unavailability, compared to 99.95%, which means 21 minutes 36 seconds unavailability per month. Multi-region also needs to be carefully considered if data governance is a concern.

Cloud SQL offers failover replica for high availability

- Replica will be created in another zone in the same region as the database
  - Uses Regional persistent disks
- Will automatically switch to the failover if the primary instance is unavailable.
- Doubles the cost of the database.



Google Cloud

The HA configuration is available for all three Cloud SQL offerings: MySQL, PostgreSQL, and SQL Server. Full details for the configuration are here: https://cloud.google.com/sql/docs/mysql/high-availability

The primary instance and standby share the same IP address so that in the event of failure, clients will seamlessly be redirected to the standby.

Spanner and Firestore can be deployed in one/multiple regions

Database	Availability SLA
Firestore single region	99.99%
Firestore multi-region	99.999%
Spanner single region	99.99%
Spanner multi-region	99.999%

Google Cloud

Both Firestore and Spanner offer single and multi-region deployment. Multi-region locations can withstand the loss of entire regions and maintain availability without losing data. However, multi-region deployments also cost more money. Which you choose ties back to your application requirements and service-level objectives.

When designing for reliability, consider these key performance metrics

Availability

The percent of time a system is running and able to process requests

- Achieved with fault tolerance.
- Create backup systems.
- Use health checks.

Durability

The odds of losing data because of a hardware or system failure

- Achieved by replicating data in multiple zones.
- Do regular backups.
- Practice restoring from backups.

Scalability

The ability of a system to continue to work as user load and data grow

- Monitor usage.
- Use capacity auto- scaling to add and remove servers in response to changes in load.

Google Cloud

Next discussion

Durability
Ensuring that data is preserved and available is a mixture of replication and backup. Regular restores from backup should be performed to confirm that the process works as expected.

Durability represents the odds of losing data

Preventing data loss is a shared responsibility.

Storage choice	Google Cloud provides	What you should do
Cloud Storage	11 9's durability Versioning (optional)	Turn versioning on
Disks	Snapshots	Schedule snapshot jobs
Cloud SQL	On-demand backups Automated backups Point-in-time recovery Failover server (optional)	Create at any time Run SQL database backups
Spanner	Automatic replication	Run export jobs
Firestore	Automatic replication	Run export jobs

Google Cloud

Durability of data represents the odds of losing the data. Depending on the storage solution, the durability is a shared responsibility. Google Cloud's responsibility is to ensure that data is durable in the event of a hardware failure. Your responsibility is performing backups of your data.

For example, Cloud Storage provides you with 11 9's durability, and versioning is a feature. However, it's your responsibility to determine when to use versioning. It is recommended that you turn versioning on and have older versions archived as part of an object lifetime management policy.

For other storage services to achieve durability, it usually means taking backups of data. For disks, this means snapshots, so snapshot jobs should be scheduled. For Cloud SQL, you can create a backup at any time (on-demand). This could be useful if you are about to perform a risky operation on your database, or if you need a backup and you do not want to wait for the backup window. Google Cloud also provides automated backups, point in time recovery, and optionally a failover server. You can create on-demand backups for any instance, whether the instance has automatic backups enabled or not. To improve durability, SQL database backups should also be run.

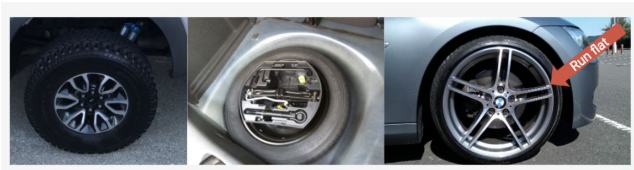
Spanner and Firestore provide automatic replication, and you should run export jobs with the data being exported to Cloud Storage.

Designing for disaster recovery...

Google Cloud

Disaster recovery patterns

- DR patterns are considered to be cold, warm, or hot.
- These patterns indicate how readily the system can recover when something goes wrong.
- An analogy might be what you would do if you were driving and punctured a car tire



Cold: You have no spare tire, so you must call someone to come to you with a new tire and replace it.

Your trip stops until help arrives to make the repair.

Warm: You have a spare tire and a replacement kit, so you can get back on the road using what you have in your car.

However, you must stop your journey to repair the problem

Hot: You have run-flat tires. You might need to slow down a little, but there is no immediate impact on your journey.

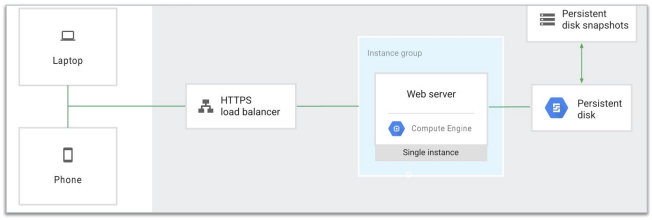
Your tires run well enough that you can continue

Google Cloud

Disaster Recovery Planning Guide
https://cloud.google.com/architecture/dr-scenarios-planning-guide

Disaster recovery: Cold standby

- One VM with an application in a MIG
  - Application stores data on attached disk
  - Snapshots are regularly taken
- Instance template has a startup script to create a persistent disk from the last disk snapshot and attach it
- A health check monitors the status of the VM inside the managed instance group.
- External HTTP(S) Load Balancing directs users to the VM that runs in the managed instance group



- If the VM fails, the managed instance group tries to recreate a VM in the same zone
  - That fails if there is a failure at the zone level
  - Manually create a replacement managed instance group in another zone (or region)
  - Edit the load balancer to point to the new MIG

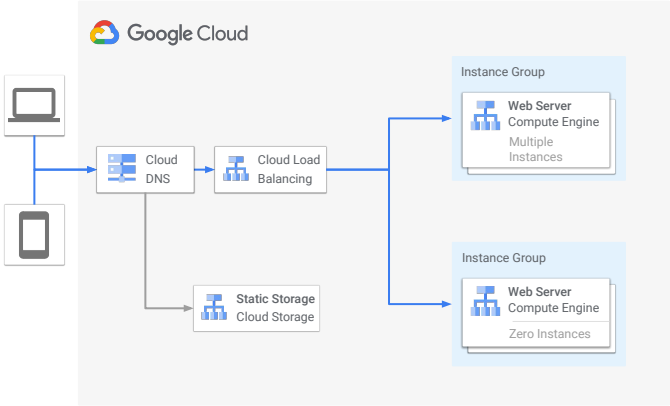
Google Cloud

From:
https://cloud.google.com/architecture/cold-recoverable-apps-persistent-disk-snapshots

- One VM runs in a managed instance group with an attached persistent disk that stores a basic website.
- Snapshots are regularly taken of the persistent disk using a resource policy snapshot schedule.
- A startup script is applied to an instance template so any VMs created in the managed instance group create a persistent disk from the last disk snapshot and attach it.
- A health check monitors the status of the VM inside the managed instance group.
- External HTTP(S) Load Balancing directs users to the VM that runs in the managed instance group.
- If the VM fails, the managed instance group tries to recreate a VM in the same zone. If the failure is at the zone level, you must manually create a replacement managed instance group in another, working zone.

Disaster recovery: warm standby

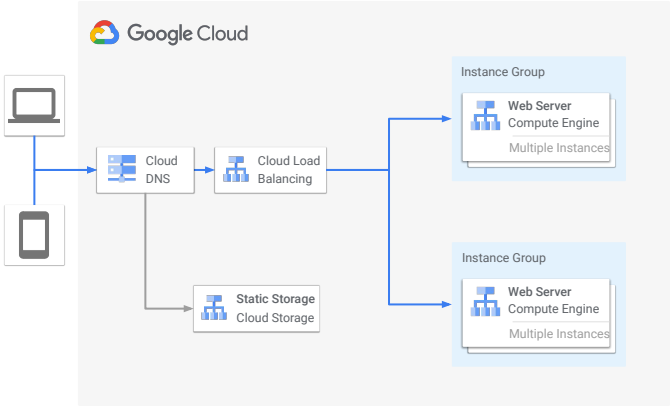
- Create instance groups in two regions
  - In one group, # of running instances = 0
- Use a global load balancer
  - All traffic is going to one region
- If region fails, spin up instances in the other instance group
  - Load balancer will automatically route traffic to it



Google Cloud

Disaster recovery: Hot standby

- Create instance groups in multiple regions.
- Use a global load balancer.
- Store unstructured data in multi-region buckets.
- For structured data, use a multi-region database such as Spanner or Firestore.



Google Cloud

The hot standby is about using Google Cloud features like instance groups deployed across multiple regions and traffic distributed using a global load balancer. Depending on the storage service, multi-region options or high availability configurations must be selected.

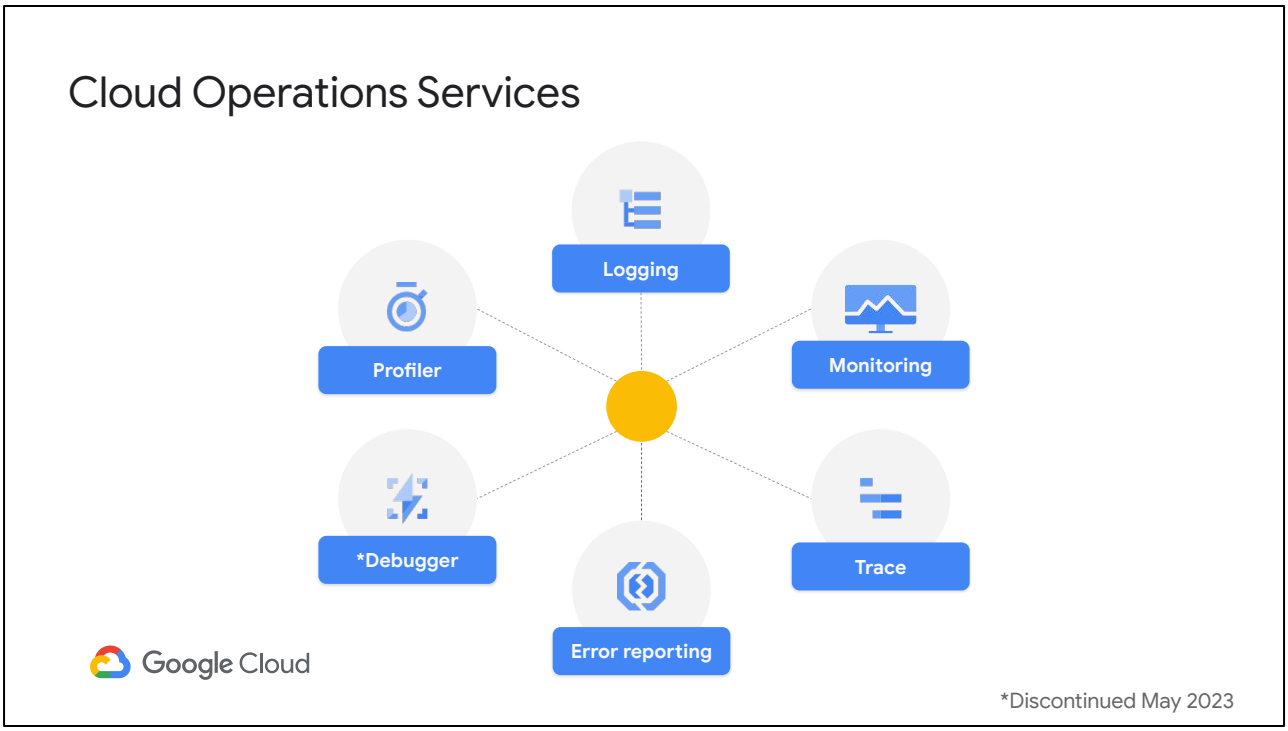
Cloud Operations...

Google Cloud

Cloud Operations Suite

- **Provides comprehensive observability of all deployed resources**
  - Integrated monitoring, logging, and trace managed services for applications and systems
  - Create and monitor service-level objectives (SLOs) as part of your SRE strategy

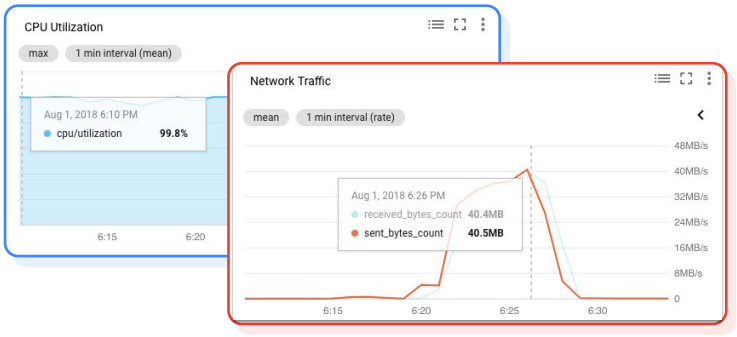
Google Cloud



- **Cloud Operations** is Google Cloud's **fully managed native logging and monitoring tool**
- **Monitoring**
  - Default monitoring for all resources, with plugins for many third party tools
  - Uptime checks for groups or specific resources
- **Logging**
  - **All logs are captured in Cloud Logging**
  - Search and filter
  - Derive metrics from logs, use to create dashboard or autoscale instances
  - Export for retention and leveraging other tools
- **Performance** - Highly scalable and performant
- **Multi-cloud** - Support Google Cloud and AWS. Can monitor on-premises using a partner solution

Cloud Monitoring dashboards can visualize utilization and network traffic

- Collects metrics, events, and metadata from
  - Google Cloud,
  - Amazon Web Services (AWS)
  - Application instrumentation
- Generates insights via dashboards, charts, and alerts
- Applications can generate custom metrics



Google Cloud

Cloud Monitoring

Cloud Monitoring
https://cloud.google.com/monitoring

Cloud Monitoring allows you to create custom dashboards that contain charts of the metrics that you want to monitor. For example, you can create charts that display your instances' CPU utilization, the packets or bytes sent and received by those instances, and the packets or bytes dropped by the firewall of those instances.

In other words, charts provide visibility into the utilization and network traffic of your VM instances, as shown on this slide. These charts can be customized with filters to remove noise, groups to reduce the number of time series, and aggregates to group multiple time series together.

For a full list of supported metrics, please refer to the documentation:
https://cloud.google.com/monitoring/api/metrics_gcp

Custom metrics

- Built-in metrics can provide information on backend latency or disk usage, for example
- Custom metrics (application-specific metrics) let you define and collect information built-in metrics cannot
  - For example, the count of number of users logged into the application
- These metrics are captured by using an API provided by a code library to instrument your code
  - Google recommends open-source OpenCensus for metric selection
    - Provides a way to create custom metrics, add metric data to those metrics, and export the metric data to Cloud Monitoring
- Can also create metrics based on the content of log entries
  - For example, number of log entries containing a particular message

Google Cloud

Custom metrics
https://cloud.google.com/monitoring/custom-metrics

OpenCensus: https://opencensus.io/

Suggested Lab: Reporting Application Metrics into Cloud Monitoring (if time allows)

- Uses OpenCensus to record metrics
- Discusses other instrumentation methods

https://partner.cloudskillsboost.google/catalog_lab/1047

Uptime checks test the availability of your public services

CHECKS	VIRGINIA	OREGON	IOWA	BELGIUM	SINGAPORE	SAO PAULO	POLICIES
Instance 1	✓	✓	✓	✓	✓	✓	
Instance 2	✓	✓	✓	✓	✓	✓	
Instance 3	✓	✓	✓	✓	✓	✓	

Google Cloud

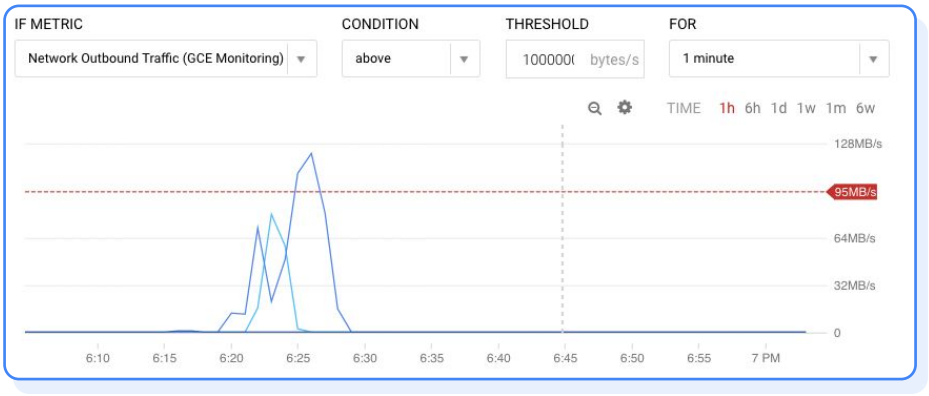
Uptime checks

Uptime checks
https://cloud.google.com/monitoring/uptime-checks

There are also uptime checks that you can configure to test the availability of your public services from locations around the world, as you can see on this slide. The type of uptime check can be set to HTTP, HTTPS, or TCP. The resource to be checked can be an App Engine application, a Compute Engine instance, a URL of a host, or an AWS instance or load balancer.

For each uptime check, you can create an alerting policy and view the latency of each global location.

Alerting policies can notify you of certain conditions



Google Cloud

Introduction to alerting

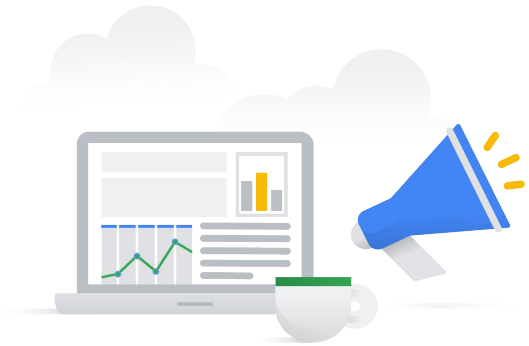
Now, although charts are extremely useful, they can only provide insight while someone is looking at them. But what if your server goes down in the middle of the night or over the weekend? Do you expect someone to always look at dashboards to determine whether your servers are available or have enough capacity or bandwidth? If not, you want to create alerting policies that notify you when specific conditions are met.

For example, as shown on this slide, you can create an alerting policy when the network egress of your VM instance goes above a certain threshold for a specific timeframe. When this condition is met, you or someone else can be automatically notified through email, SMS, or other channels in order to troubleshoot this issue.

Alerting Policies

Create notifications in response to a policy that exceeds some condition

1. Create a condition that determines when some metric exceeds some value for a specified period of time.
2. Specify a notification – can be a text, email, webhook, and others.
3. Can add additional documentation to the notification.
4. Name the policy.



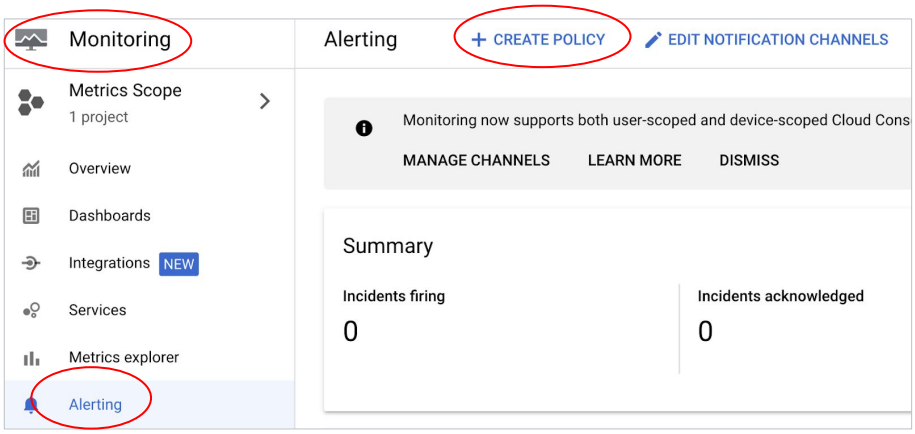
Google Cloud

An alerting policy allows you to respond to anomalies in your system. For example, if your system starts to get an unusually high number of requests, you might want to be notified of a potential denial of service attack. Or, if a VM is not working, you can have the system recreate it.

To create an alert, you define a condition that determines if some metric is above or below some value for some period of time.

When the condition is met, you respond with a notification. Notifications include emails, texts, and webhooks. Use an email or text to notify a human. Use a webhook to run a program in response to the anomaly.

Creating Alerts




```
gcloud monitoring policies create
--policy-from-file="rising-cpu-usage.json"
```

Google Cloud

Alerting Policies Illustrated

1 **Conditions**

Basic Conditions

Suggested Condition — Edit Delete

Violates when: Any appengine.googleapis.com/system/instance_count stream is above a threshold of 10 for greater than 1 minute

+ Add Another Condition

2 **Notifications** (optional)

When alerting policy violations occur, you will be notified via these channels. Learn more

Email — doug@roitraining.com

+ Add Another Notification

3 **Documentation** (optional)

When email notifications are sent, they'll include any text entered here. This can convey useful information about the problem and ways to approach fixing it.

Edit Preview — Markdown Formatting Help

Hey, here's some documentation of this alert!

4 **Name this policy**

A policy's name is used in identifying which policies were triggered, as well as managing configurations of different policies.

Threshold - user/dougs-cf-log-metric

Save Policy Cancel

Google Cloud

You can include additional information as needed along with the alert. Creating the alert can be done right from the Web Console if you like.

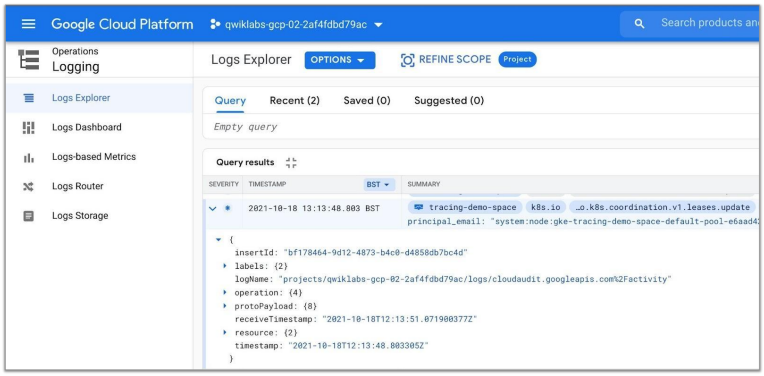
Suggested Lab: Monitoring & Alerting (if time allows)

https://partner.cloudskillsboost.google/catalog_lab/755

Google Cloud

Cloud Logging collects logs from admin, system and application activity

- Data is available from over 150 common application components, on-premises systems, and hybrid cloud systems
- Store, search, analyze, monitor, and alert on logging data from Google Cloud and Amazon Web Services



[Cloud Logging pricing for Cloud Admins: How to approach it & save cost]

Google Cloud

Cloud Logging pricing for Cloud Admins: How to approach it & save cost
https://cloud.google.com/blog/topics/cost-management/how-to-approach-cloud-logging-pricing-for-cloud-admins

Cloud Logging
https://cloud.google.com/logging/docs

Log data compliance:
https://services.google.com/fh/files/misc/whitepaper_data_governance_logs_how_to.pdf

Monitoring & Logging Agents

- Automatically stream logs and metrics from third-party applications, system logs, and your code to the Operations logs and Monitoring
- Install on VMs running Linux or Windows
  - AWS EC2 VMs also supported
  - Not required for App Engine or Kubernetes clusters

Ops Agent Overview

Google Cloud

Monitoring and Logging agents
Legacy method: Installing the Cloud Logging agent on individual VMs:
https://cloud.google.com/logging/docs/agent/logging/installation

Legacy method: Installing the Cloud Monitoring agent on individual VMs:
https://cloud.google.com/monitoring/agent/monitoring/installation

Currently recommended method: Ops Agent overview:
https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent

If you're using Kubernetes, Cloud Functions, or App Engine to deploy your applications, then all messages sent to standard out end up in the Operations logs automatically. When using virtual machines though, you will need to add the Logging Agent when you create the VM.

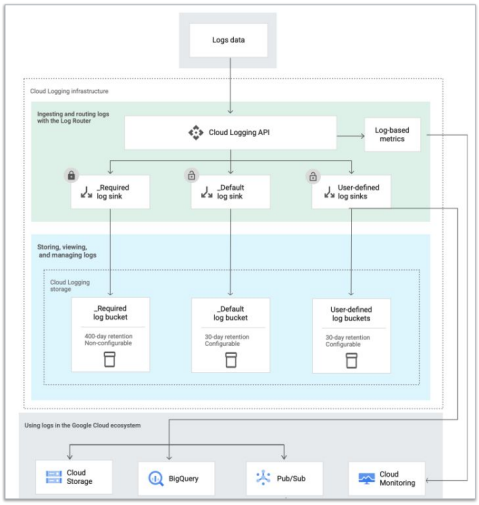
Once the Logging Agent is added, any messages you log from your application will go to Operations. Also, the logs for third-party applications like Apache web server will be aggregated into the Operations logs.

The Logging Agent can also be used on VMs running in AWS EC2.

Just add a couple lines of code to your startup script.

See the documentation for details.

High level look at log routing



Google Cloud

Routing and storage overview

https://cloud.google.com/logging/docs/routing/overview

Log Sinks

Provides a method to retain logs for longer periods of time or to stream logs to other applications.

- To export specific logs write a filter that selects them, and choose a destination in Cloud Storage, BigQuery, Pub/Sub, Cloud Logging or Splunk
  - The filter and destination are held in an object called a sink
    - Can be created at the organisation, folder, project, and billing account level
- A sink can only export logs that belong to its parent resource, e.g. project/folder.
  - When a log comes in that matches a filter, a copy of the log is written to the export destination.

Google Cloud

Sinks

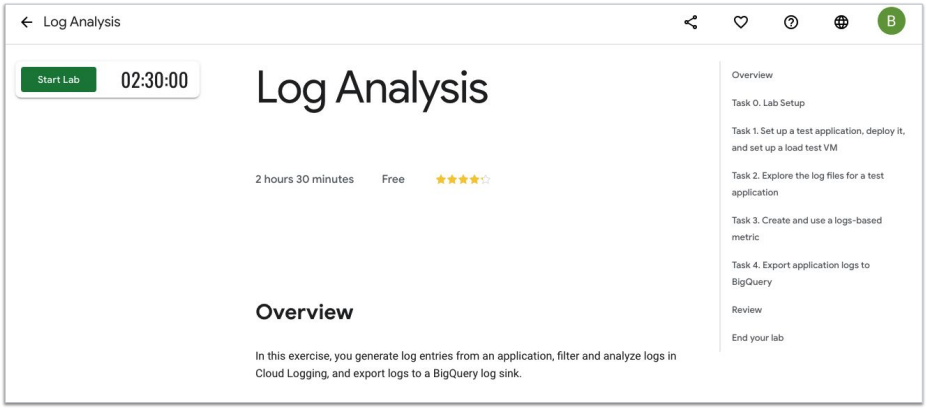
Configuring log sinks to export logs to external systems (e.g., on-premises or BigQuery)
https://cloud.google.com/logging/docs/routing/overview

Scenarios for exporting Cloud Logging data - Splunk
https://cloud.google.com/architecture/exporting-stackdriver-logging-for-splunk

View logs in sink destinations:
https://cloud.google.com/logging/docs/export/using_exported_logs

Log sinks:
https://cloud.google.com/logging/docs/export/configure_export_v2

Suggested Lab: Log Analytics (if time allows)

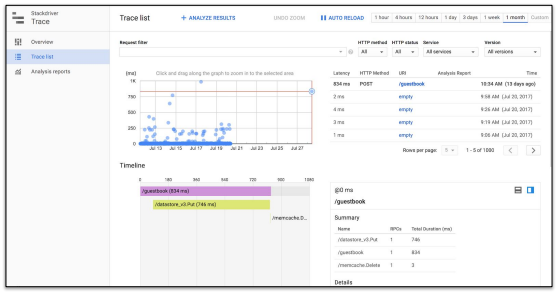


https://partner.cloudskillsboost.google/catalog_lab/2649

Google Cloud

Cloud Trace

- Collects latency data from applications
- Useful for finding performance bottlenecks and detecting issues in near-real time
- Displays requests along with their timings
- Provides information on
  - How long it takes applications to handle incoming requests from users or other applications
  - How long it takes to complete operations like RPC calls performed when handling the requests



Google Cloud

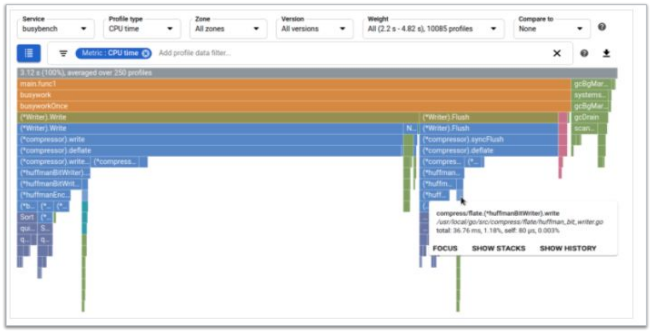
Viewing trace details

Cloud Trace
https://cloud.google.com/trace

Viewing trace details:
https://cloud.google.com/trace/docs/viewing-details

Cloud Profiler

- Continuously gathers CPU usage and memory-allocation information from production applications
  - Attributes that information to the application's source code
  - Helps identify the parts of the application consuming the most resources

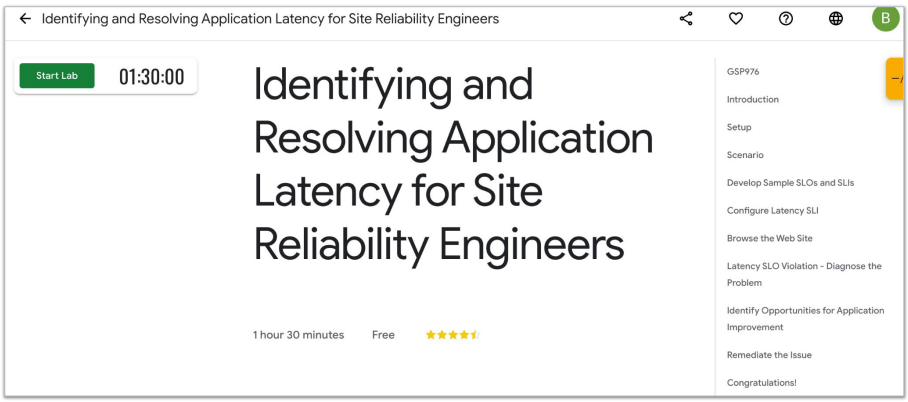


Google Cloud

Cloud Profiler

Cloud Profiler
https://cloud.google.com/profiler/docs

Suggested Lab: Trace and Profiler (if time allows)



https://partner.cloudskillsboost.google/catalog_lab/4138

Google Cloud

Cloud Error Reporting

- A single place to monitor error conditions from all apps and services in a Google Cloud project and from Amazon Elastic Compute Cloud (EC2) applications
- Counts, analyzes, and aggregates the crashes in your running cloud service
- Opt-in to receive real-time alerts

Google Cloud

Cloud Error Reporting

Cloud Error Reporting
https://cloud.google.com/error-reporting

Suggested Lab: Error Reporting (if time allows)

https://partner.cloudskillsboost.google/catalog_lab/4357

Google Cloud

Google Cloud