

Welcome to Google Cloud

We're thrilled to have you join the Google Cloud Startups ecosystem.

Your account team is here to champion your success and help you navigate the resources and support available.

Google Cloud



Onboarding guide content

REQUIRED

1. Set up your Google Cloud Organization
2. Setup IAM permissions
3. Purchase Support
4. Enable Security Command Center

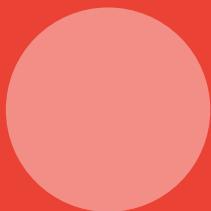
5. Set up Budgets and Alerts
6. Visualize your spending with Looker Studio

RECOMMENDED

INFORMATIONAL

7. Google for Startups program journey
8. Request free on-demand training licenses
9. Access the Gemini APIs
10. Use AI Hypercomputers
11. File a support case
12. Next steps

Set up your Google Cloud Organization



GCP | Set up your organization

You can create a Google Cloud organization by purchasing a Google Workspace subscription or by creating an organization directly with Cloud Identity (without Google Workspace):

Google Workspace

[Sign up for Google Workspace.](#)

Google for Startups program provides a free 1YR Google Workspace Business Plus subscription for new Google Workspace customers

[Compare Google Workspace editions](#)

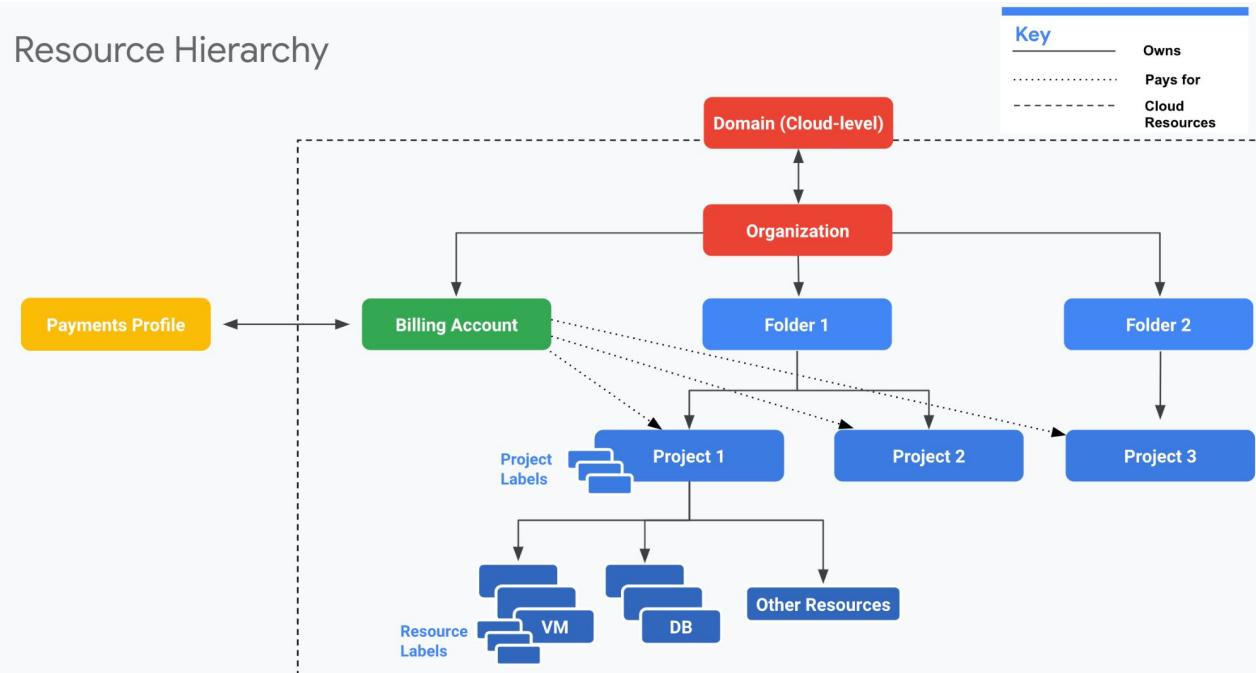


[Sign up for Cloud Identity.](#)

Cost Effective
Free tier provides 50 user licenses by default

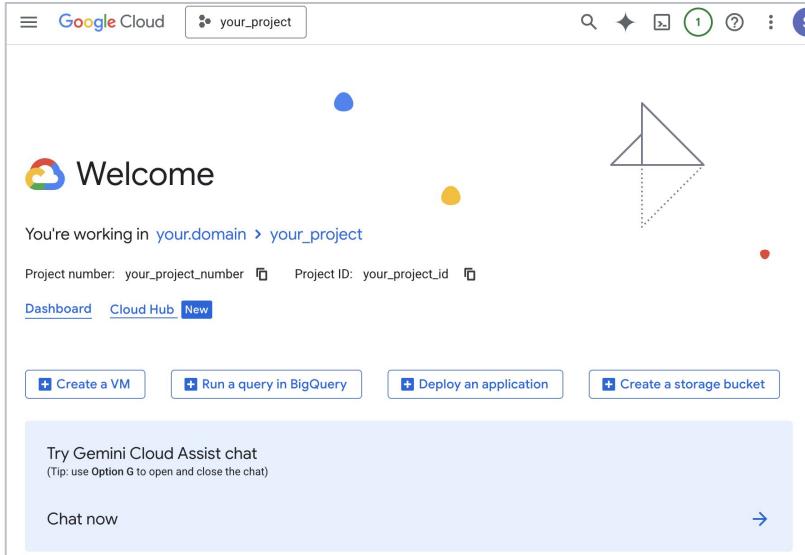
GCP | Setup additional resources

Please follow our [guided flow](#) to set up additional resources in your organization



GCP | Consoles

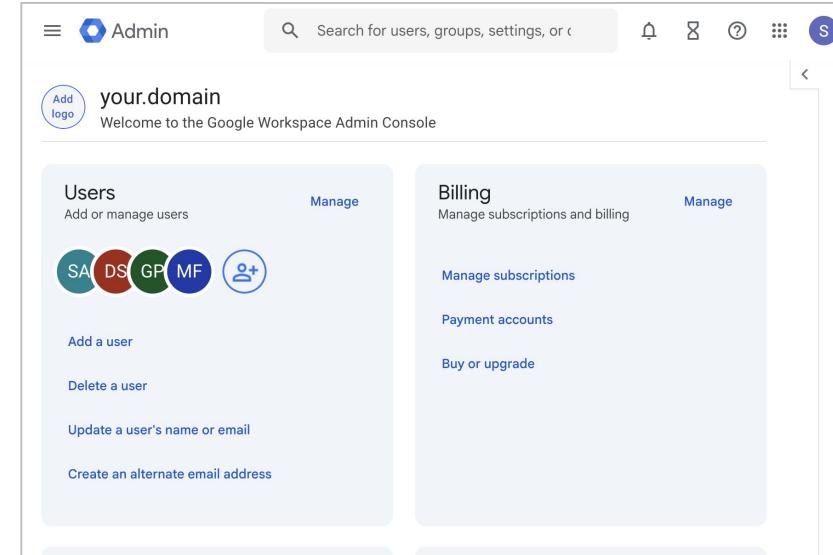
console.cloud.google.com



The screenshot shows the Google Cloud Console homepage. At the top, there's a navigation bar with 'Google Cloud' and a dropdown for 'your_project'. Below the navigation is a large 'Welcome' section with a colorful geometric graphic. It displays the message 'You're working in your.domain > your_project'. Below this, it shows 'Project number: your_project_number' and 'Project ID: your_project_id'. There are three main navigation buttons: 'Dashboard', 'Cloud Hub', and 'New'. Under the 'New' button, there are four buttons: '+ Create a VM', '+ Run a query in BigQuery', '+ Deploy an application', and '+ Create a storage bucket'. At the bottom, there's a 'Try Gemini Cloud Assist chat' section with a 'Chat now' button.

Allows you to implement granular access controls via Identity and Access Management (IAM), enable APIs and services, create and manage infrastructure and much more.

admin.google.com



The screenshot shows the Google Workspace Admin Console homepage. At the top, there's a navigation bar with 'Admin' and a search bar. Below the navigation is a section for 'your.domain' with the message 'Welcome to the Google Workspace Admin Console'. There are two main sections: 'Users' and 'Billing'. The 'Users' section includes a 'Manage' button and a grid of user icons (SA, DS, GP, MF). It also has buttons for 'Add a user', 'Delete a user', 'Update a user's name or email', and 'Create an alternate email address'. The 'Billing' section includes a 'Manage' button and buttons for 'Manage subscriptions', 'Payment accounts', and 'Buy or upgrade'.

Allows you to manage people, groups, access controls, and Google Workspace domains

Setup your Google Cloud Organization



Please use our [Signup Issues Toolkit](#) to:

1. Recover your Google Admin (admin.google.com) account
2. Fix the “Domain already in use” issue



Additional tasks

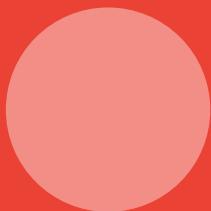
1. [Add an additional Super Admin to your organization](#)
2. [Migrating projects between organization resources](#)
3. [Enable, disable, or change billing for a project](#)
4. [Creating and managing organization policies](#)



Best practices:

1. [Super administrator account best practices](#)
2. [Security best practices for administrator accounts](#)

Setup IAM permissions



IAM | Get to the organization view

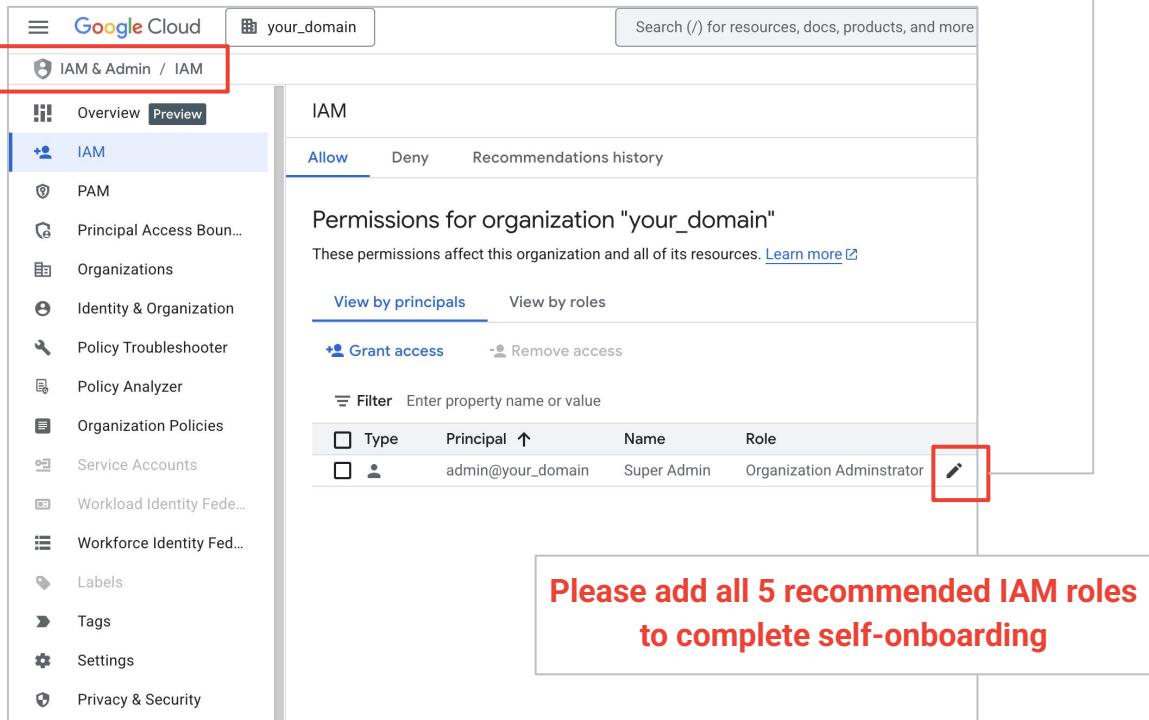
The resource selector is at the top left side of your screen

The image shows a screenshot of the Google Cloud Resource Selector. On the left, the selector displays the project name 'project_name' with a red box around it. On the right, a 'Select a resource' dialog is open, showing a dropdown menu with 'google.com' and a search bar. Below the search bar, there are tabs for 'Recent', 'Starred', and 'All'. A table lists resources: 'project_name' (Project, ID: project_id, star icon) and 'google.com' (Organization, ID: organization_id, star icon). A red box highlights the 'google.com' row. At the bottom of the dialog, the text 'select your domain/organization' is displayed.

Name	Type	ID
✓ project_name ⓘ	Project	project_id
google.com ⓘ	Organization	organization_id

select your domain/organization

IAM | Add recommended roles



Google Cloud your_domain Search (/) for resources, docs, products, and more

IAM & Admin / IAM

Overview Preview

IAM

Allow Deny Recommendations history

Permissions for organization "your_domain"

These permissions affect this organization and all of its resources. [Learn more](#)

View by principals View by roles

Grant access Remove access

Filter Enter property name or value

Type	Principal ↑	Name	Role
<input type="checkbox"/>	<input type="checkbox"/> admin@your_domain	Super Admin	Organization Administrator

Please add all 5 recommended IAM roles to complete self-onboarding

Edit access to "your_domain"

Principal [?](#) Organization
admin@your_domain your_domain

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. [Learn more](#)

IAM condition (optional) [?](#)
+ Add IAM condition

Role Organization Administrator [▼](#)

Access to manage IAM policies and view organization policies for organizations, folders, and projects.

IAM condition (optional) [?](#)
+ Add IAM condition

Role Billing Account Administrator [▼](#)

Authorized to see and manage all aspects of billing accounts.

IAM condition (optional) [?](#)
+ Add IAM condition

Role Support Account Administrator [▼](#)

Allows management of a support account without giving access to support cases.

IAM condition (optional) [?](#)
+ Add IAM condition

Role Tech Support Editor [▼](#)

Full read-write access to technical support cases (applicable for GCP Customer Care and Maps support).

IAM condition (optional) [?](#)
+ Add IAM condition

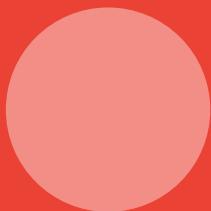
Role Security Center Admin [▼](#)

Admin(super user) access to security center

+ Add another role

Save Test changes [?](#) Cancel

Purchase Support



Your account team **does not** engage in troubleshooting.

Basic Support (default) **does not** give you access to our support team.

Having a paid support subscription, will ensure that your issues in production are resolved in a timely manner.

Google for Startups - Scale Tier or above already provides dedicated credits to purchase Enhanced Support

Support | Customer Care Portfolio

If you received \$12k in support credits, we **REQUIRE** you to purchase **Enhanced Support**

You **WILL NOT** be able to file support cases if you are in the **Basic Support tier** (Free / default)

	Basic Support	Standard Support	Enhanced Support	Premium Support
Documentation 🔗	✓	✓	✓	✓
Community Support 🔗	✓	✓	✓	✓
Multi-channel billing support 🔗	✓	✓	✓	✓
Unlimited access to support ⓘ		✓	✓	✓
Multi-channel technical support		✓	✓	✓
Available languages 🔗		English	English, Japanese, Mandarin, Chinese, and Korean	English, Japanese, Mandarin, Chinese, and Korean
Service times 🔗		8/5 response for high-impact issues	24/7 response for high & critical-impact issues	24/7 response for high & critical-impact issues
Priority issues/cases 🔗		4-hour response time	1-hour response time	15-minute initial response time
Active Assist Recommendations API ⓘ		✓	✓	✓
Issue/case escalations ⓘ			✓	✓

Standard Support is recommended if you did not receive support credits

Support | Navigate to the Support page

console.cloud.google.com/support

The screenshot shows the Google Cloud Support page. A red box highlights the 'Overview' tab in the navigation menu on the left. Another red box highlights the 'View Customer Care services' link in the 'Support information' section.

Google Cloud Support Overview

This information applies only to support for Google Cloud Platform products. You can find support options for other products, such as Maps and Apps APIs, in their own developer documentation site. Go to [Products](#) for a complete list.

Google Cloud incidents

No project selected
Please select a project to view incidents

→ Go to Personalized Service Health

Support information

Your current Customer Care service: Standard

Get responses within four business hours for P2 cases

Resource: Check out actionable recommendations to keep your cloud optimized

→ View Customer Care services

→ Manage IAM permissions

→ Get started with Standard Support

?

Get help

Support | Purchase a support subscription



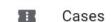
Support



Customer Care services



Overview



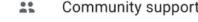
Cases



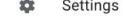
Billing support



Phone support



Community support



Settings

Sign up for Customer Care

Get the most from Google Cloud with Customer Care, choose the technical support service that's right for your organization, and further tailor it to fit your needs with additional Value-Add Services.

Basic Support

\$0

Standard Support

Minimum spend of \$29.00
OR
3% of [monthly Cloud charges](#)

Charges will be at least the minimum charge of \$29.00 or the result of the calculation, whichever is higher.

[Calculate cost](#)

Enhanced Support

Minimum spend of \$100.00
OR
10% of [monthly Cloud charges](#) for the first \$0-\$10K

7% of [monthly Cloud charges](#) from \$10K-\$80K

5% of [monthly Cloud charges](#) from \$80K-\$250K

3% of [monthly Cloud charges](#) over \$250K

Charges will be at least the minimum charge of \$100.00 or the result of the calculation, whichever is higher.

[Calculate cost](#)

Premium Support

Minimum spend of \$15,000.00
OR
10% of [monthly charges](#) for the first \$0-\$150K

7% of [monthly charges](#) from \$150K-\$500K

5% of [monthly charges](#) from \$500K-\$1M

3% of [monthly charges](#) over \$1M

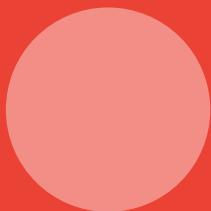
Charges will be at least the minimum charge of \$15,000.00 or the result of the calculation, whichever is higher.

[Calculate cost](#)

If "BUY NOW" is grayed out, review your IAM permissions again.

[BUY NOW](#)
[BUY NOW](#)
[CONTACT SALES](#)

Enable Security Command Center



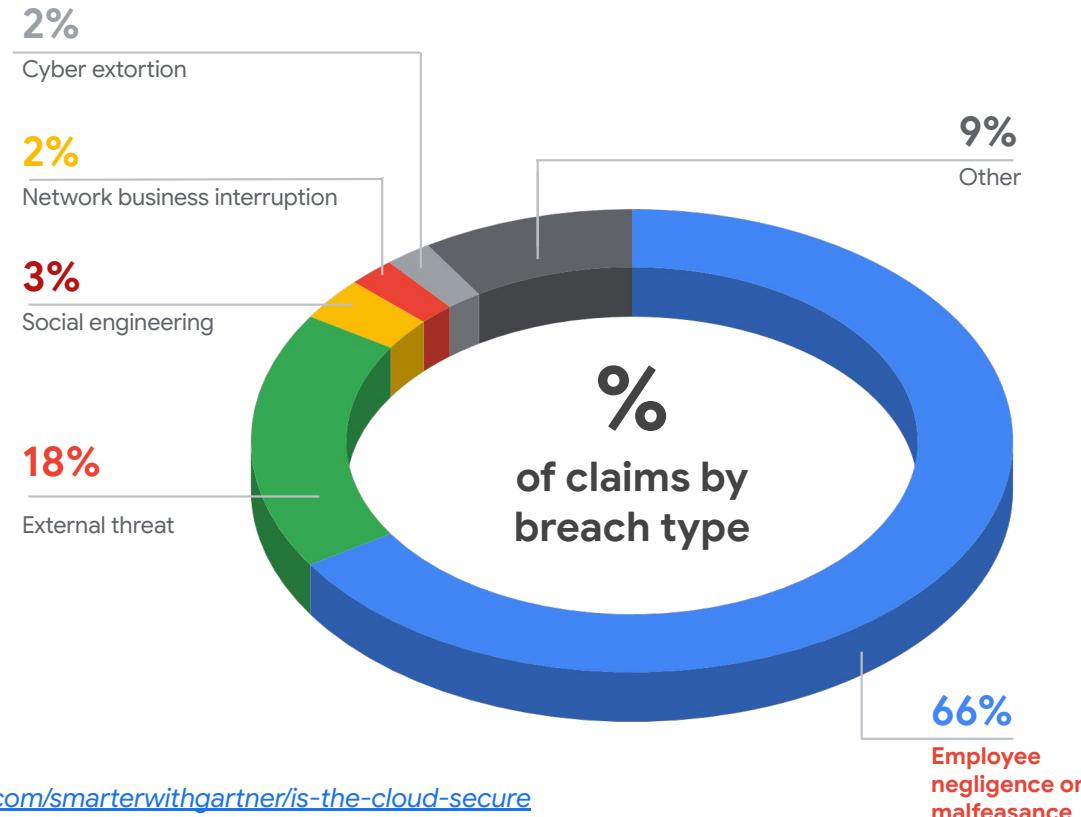
SCC | Security is a shared responsibility.

You're the expert in knowing the security and regulatory requirements for your business, and knowing the requirements for protecting your confidential data and resources. When you run your workloads on Google Cloud, you must identify the security controls that you need to configure in Google Cloud to help protect your confidential data and each workload. To decide which security controls to implement, you must consider the following factors:

- Your regulatory compliance obligations
- Your organization's security standards and risk management plan
- Security requirements of your customers and your vendors

If you do not have a security engineer, we highly recommend that you purchase SCC Premium to reduce misconfigurations and understand your security posture.

SCC | Misconfigurations are the largest cause of breaches



To learn more: gartner.com/smarterwithgartner/is-the-cloud-secure

Google Cloud

SCC | Some misconfiguration surfaced by SCC Premium

Storage



- Publicly exposed buckets
- Use of legacy bucket ACLs

Networking



- Overly permissive firewall rules
- Use of default and/or legacy networks
- Subnetworks that do not use private access to Google APIs

Logging/ Monitoring



- Monitoring disabled
- Storage buckets with logging disabled
- Stackdriver monitoring for Kubernetes clusters not enabled
- VPC Flow logs disabled

VM Instances



- IP forwarding enabled
- SSH and access misconfigurations

GKE Clusters



- Private cluster disabled
- Network policy disabled
- Master authorized network disabled
- IP alias disabled
- Legacy authorization enabled

Compliance



- Monitoring against the CIS GCP Foundation benchmark
- Certified for CIS Benchmarks 1.0 and 1.1, with more certifications to come
- PCI, ISO, and NIST support

SCC | Navigate to the SCC page

Google Cloud your_domain

Security Command Center

Search Results

Risk Overview

Security

Showing resource results for `your_domain` only

Get Started Today

- 1 Select Premium Tier for the best security
- 2 Configure detectors and services for the right

Visibility

Risk Management

Threat Detection

Compliance Reporting

Asset Inventory

Platform Misconfigurations

OS, Application, Web app vulnerabilities

Malicious activity in your Cloud Platform

Malicious activity in your computer

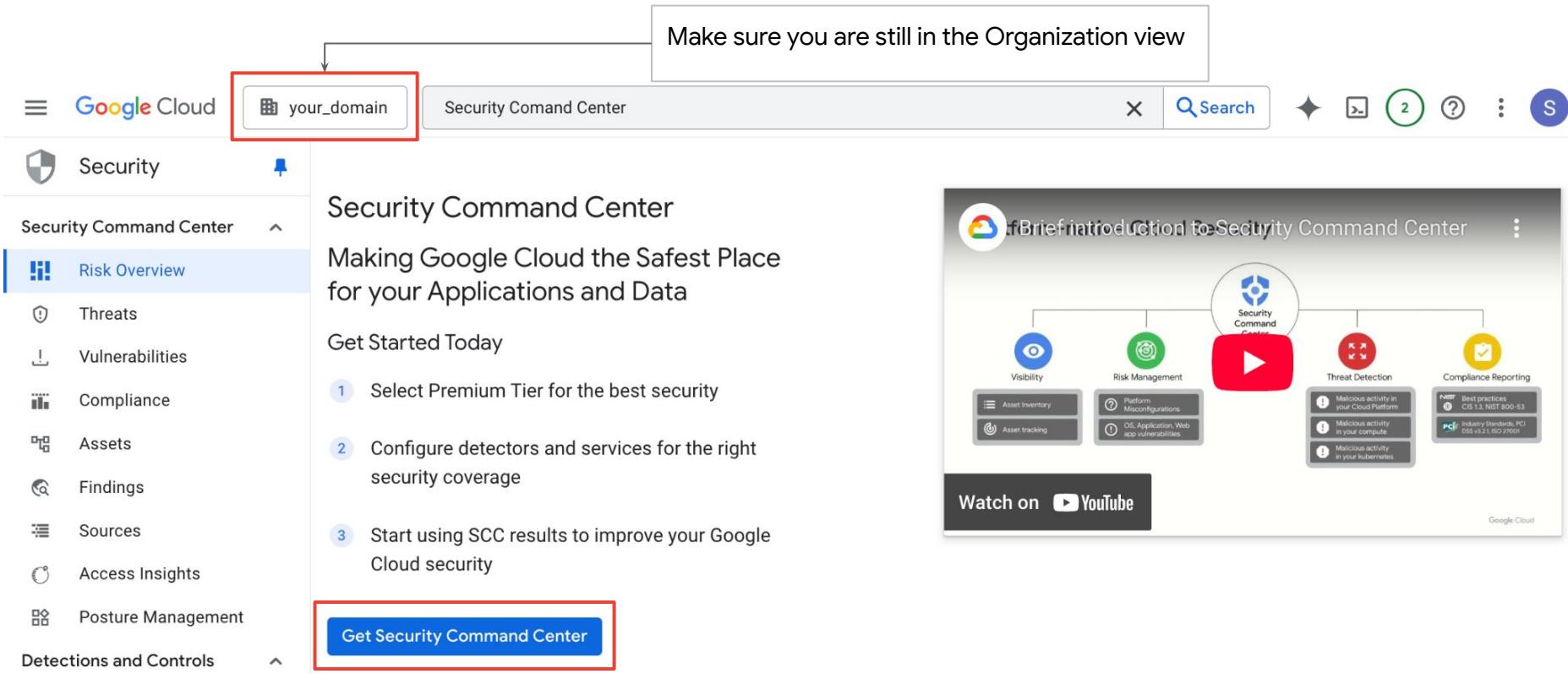
Malicious activity in your Kubernetes

PCI DSS, HIPAA, ISO 27001

GDPR, CCPA, NIST 800-53

SCC | Navigate to the SCC page

Make sure you are still in the Organization view



Google Cloud

your_domain

Security Command Center

Search

2

?

S

Security

Security Command Center

Risk Overview

Threats

Vulnerabilities

Compliance

Assets

Findings

Sources

Access Insights

Posture Management

Get Started Today

- 1 Select Premium Tier for the best security
- 2 Configure detectors and services for the right security coverage
- 3 Start using SCC results to improve your Google Cloud security

Get Security Command Center

Brief introduction to Security Command Center

Watch on YouTube

Google Cloud

SCC | Select SCC Premium

Google Cloud your_domain Security Command Center X Search

Security Command Center

Get Security Command Center

Security Command Center Risk Overview

- Threats
- Vulnerabilities
- Compliance
- Assets
- Findings
- Sources
- Access Insights
- Posture Management

Detections and Controls

- Google SecOps
- reCAPTCHA
- Model Armor
- Web Security Scanner
- Cyber Insurance Hub
- Binary Authorization
- Advisory Notifications
- Access Approval
- Managed Microsoft AD

Data Protection

- Sensitive Data Protection
- Data Loss Prevention

Get started

1 Select a tier

2 Select services

3 Grant roles

4 Complete setup

Enterprise

Includes cloud-native application protection capabilities for your organization, such as asset discovery, enhanced threat detection, case management, and more for your Google Cloud or multicloud estate.

Contact us

Premium

Includes enhanced vulnerability scanning and detection of misconfigurations. PLUS advanced Google Cloud-native Threat detection, Compliance reporting for regulatory and industry mandates, and more.

Selected

Standard

Includes vulnerability scanning for certain application security risks and detection of basic misconfigurations that could weaken your security posture.

Select

Pricing

Based on the usage of certain Google Cloud and AWS services. [View pricing details](#)

Based on the usage of certain Google Cloud services. [View pricing details](#)

Or, get a subscription for your organization. [Contact us](#)

Free

What you get

- ✓ External asset discovery with Mandiant Attack Surface Management
- ✓ Google Security Operations with enhanced threat detection and case management
- ✓ Enhanced compliance reporting
- ✓ Multicloud findings
- ✓ Enhanced vulnerability scanning
- ✓ Misconfiguration identification
- ✓ Threat detection
- ✓ Compliance reporting
- ✓ Asset query capability
- ✓ Application vulnerability scanning
- ✓ Basic misconfiguration identification

Learn more

[Security Command Center Enterprise](#)

[Security Command Center Premium](#)

[Security Command Center Standard](#)

Next

Select Premium, click “Next”

SCC | Enable premium services

Available for Premium or Enterprise
Use kernel-level instrumentation to identify potential compromise of GKE containers, including suspicious binaries.
[Learn more about Container Threat Detection](#)

Container Threat Detection deploys a DaemonSet on your clusters. Ensure your clusters have sufficient resources available to run these DaemonSets. [Learn more](#)

Virtual Machine Threat Detection

Available for Premium or Enterprise
Analyze Compute Engine instances to identify threats, including cryptomining abuse. [Learn more about Virtual Machine Threat Detection](#)

Supported cloud providers [?](#)
Google Cloud, AWS (Preview)

Cloud Run Threat Detection [Preview](#)
Available for Premium or Enterprise
Analyze Cloud Run workloads to identify threats, including malicious binary, reverse shell and others. [Learn more](#)

Vulnerability Assessment

Available for Premium or Enterprise
Scan your Compute Engine instances for common vulnerabilities and exposures (CVE).
Supported cloud providers [?](#)
Google Cloud (Preview), AWS

Notebook Security Scanner [Preview](#)
Available for Premium or Enterprise
Scan your Colab Enterprise notebooks for package vulnerabilities in the integrated open-source Python packages. [Learn more](#)

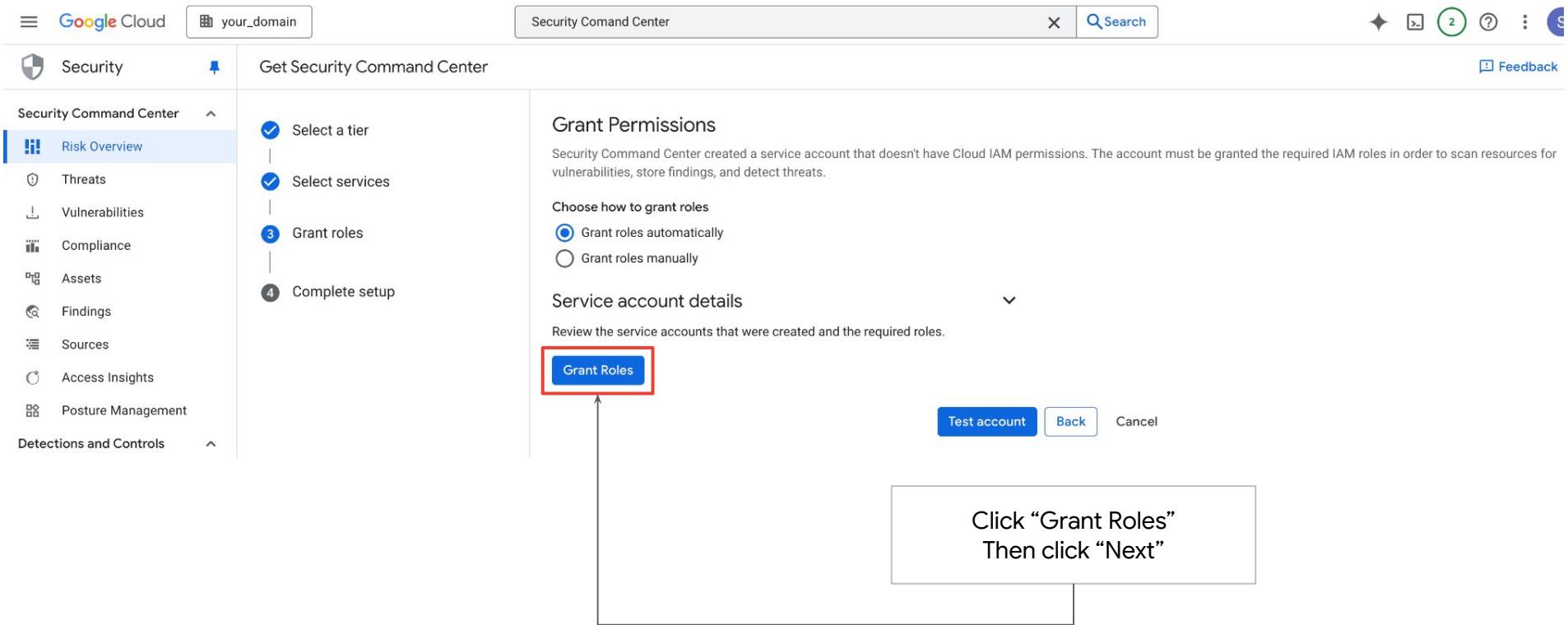
Enable [Disable](#)

Enable Cloud Run Threat Detection if necessary

Click “Next”

Next **Back** **Cancel**

SCC | Grant Permissions



Google Cloud your_domain Security Command Center Search

Security

Get Security Command Center

Feedback

Security Command Center

Risk Overview

- Threats
- Vulnerabilities
- Compliance
- Assets
- Findings
- Sources
- Access Insights
- Posture Management

Detections and Controls

Grant Permissions

Security Command Center created a service account that doesn't have Cloud IAM permissions. The account must be granted the required IAM roles in order to scan resources for vulnerabilities, store findings, and detect threats.

Choose how to grant roles

Grant roles automatically

Grant roles manually

Service account details

Review the service accounts that were created and the required roles.

Grant Roles

Test account Back Cancel

Click "Grant Roles" Then click "Next"

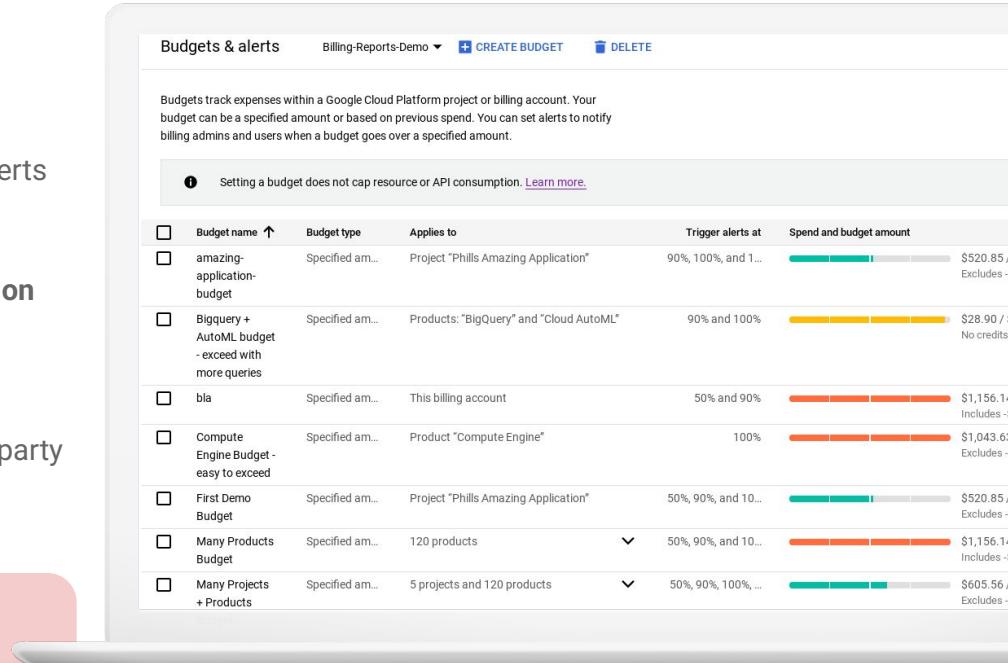
Google Cloud

Set up Budgets and Alerts

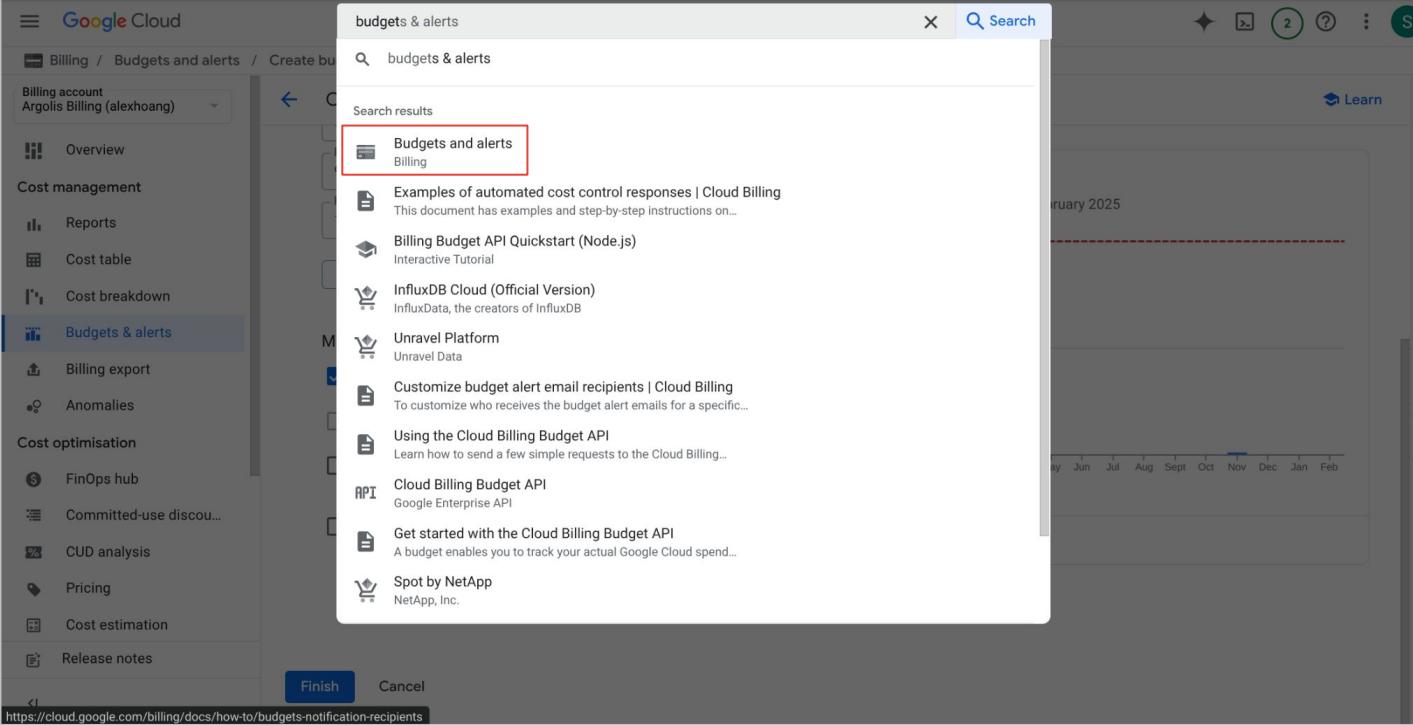


Budget Alert | Billing budgets and alerts

- A budget can be set for a project to send alerts
- Budgets may be set for a **billing account** or for a **project**. After the monthly budget is set, custom thresholds for alerts may be set (for example, 50%, 75%, 100%)
- **Reaching a budget** or threshold has **no resource restriction implications**. All GCP resources continue to function normally.
- You can use Cloud Billing budget notifications with third-party or homegrown cost-management solutions, as well as Google Cloud services.
- **We do not process refunds for runaway bills**, so please set up a budget alert to avoid them.

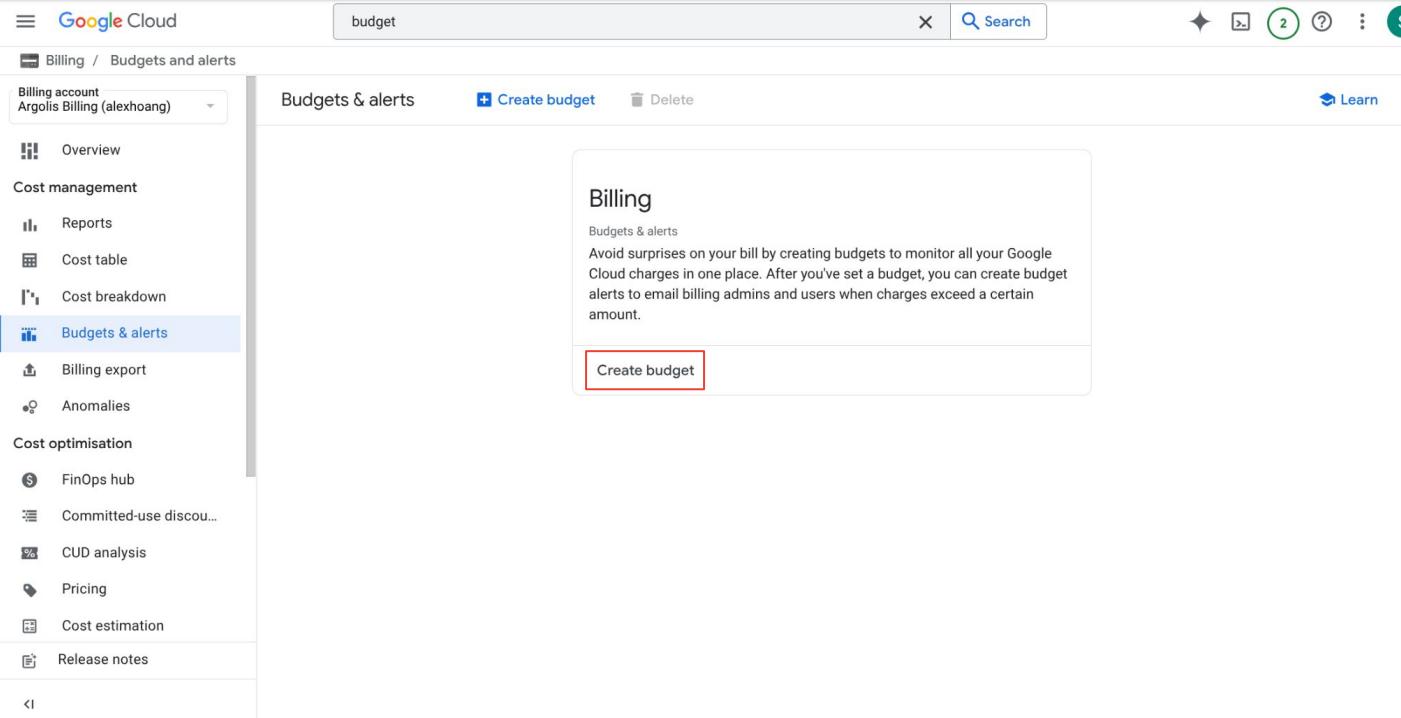


Budget Alert | Navigating the console



The screenshot shows the Google Cloud Billing interface. The left sidebar is titled 'Billing' and includes sections for 'Overview', 'Cost management' (Reports, Cost table, Cost breakdown, Budgets & alerts), 'Cost optimisation' (FinOps hub, Committed-use discounts, CUD analysis, Pricing, Cost estimation, Release notes), and 'API' (Cloud Billing Budget API, Get started with the Cloud Billing Budget API, Spot by NetApp). The 'Budgets & alerts' section is currently selected. The main content area shows a search results page for 'budgets & alerts'. The first result, 'Budgets and alerts' under 'Billing', is highlighted with a red box. Other results include 'Examples of automated cost control responses | Cloud Billing', 'Billing Budget API Quickstart (Node.js)', 'InfluxDB Cloud (Official Version)', 'Unravel Platform', 'Customize budget alert email recipients | Cloud Billing', 'Using the Cloud Billing Budget API', 'Cloud Billing Budget API', 'Get started with the Cloud Billing Budget API', and 'Spot by NetApp'. At the bottom of the search results, there are 'Finish' and 'Cancel' buttons. The URL in the address bar is <https://cloud.google.com/billing/docs/how-to/budgets-notification-recipients>.

Budget Alert | Create



The screenshot shows the Google Cloud Billing & alerts interface. The left sidebar is titled "Billing account" and shows "Argolis Billing (alexhoang)". The "Cost management" section is expanded, with "Budgets & alerts" selected and highlighted with a blue background. The main content area is titled "Billing" and contains a "Create budget" button, which is also highlighted with a red box. The "Create budget" button is located within a box that also contains the text: "Avoid surprises on your bill by creating budgets to monitor all your Google Cloud charges in one place. After you've set a budget, you can create budget alerts to email billing admins and users when charges exceed a certain amount." The top navigation bar includes a search bar with the word "budget", a "Search" button, and a "2" icon indicating two notifications.

Budget Alert | Define settings

Define the scope:

- Time range
- Projects, services, labels

Set the budget amount:

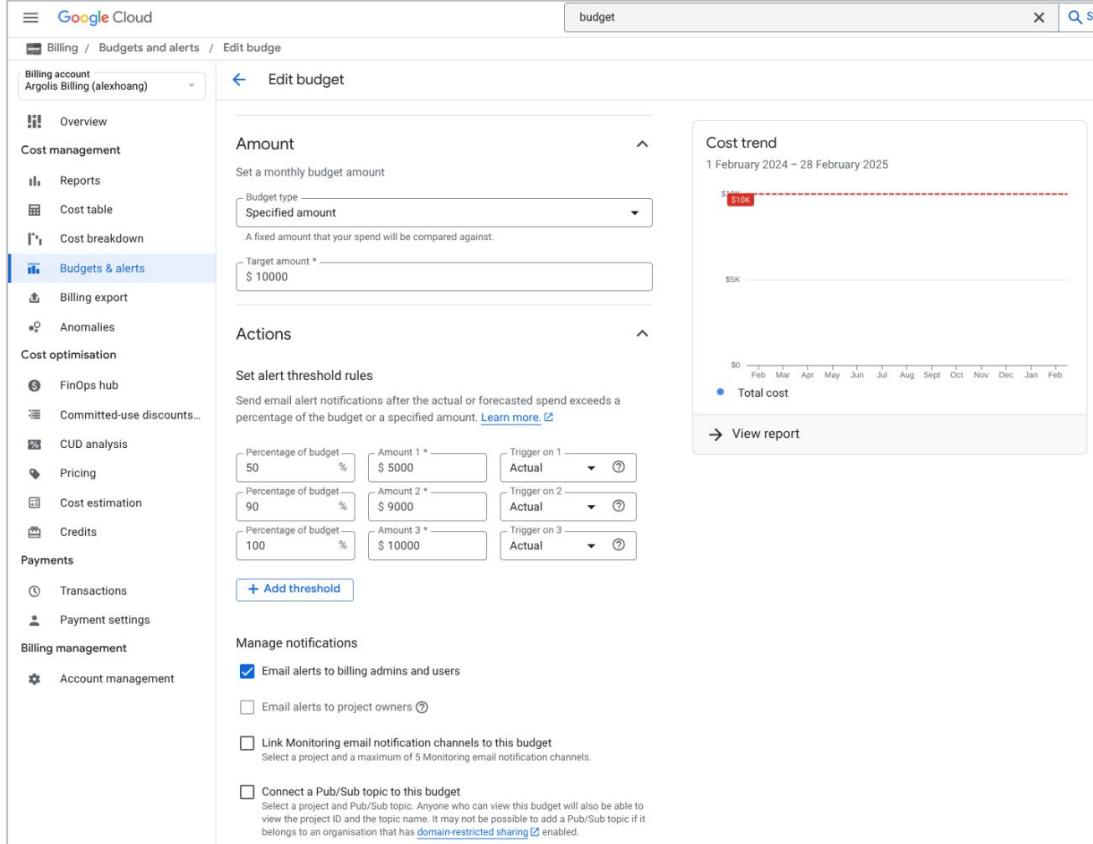
- Specified amount or last period's spend

Configure alerts:

- Thresholds (%), amount)
- Trigger on actual or forecasted spend

Manage notifications:

- Email to billing admins, project owners, or custom recipients
- Programmatic notifications with Pub/Sub



The screenshot shows the Google Cloud Billing & Budgets and alerts interface. The left sidebar lists various cost management and optimization tools. The main 'Edit budget' page is displayed, showing the 'Amount' section where a monthly budget of \$10,000 is set. The 'Actions' section contains 'Set alert threshold rules' for email notifications. Three alert rules are defined, each with a percentage of budget (50, 90, 100) and an amount (\$5000, \$9000, \$10000), and a trigger condition set to 'Actual'. Below this, 'Manage notifications' includes checkboxes for email alerts to billing admins, project owners, and monitoring channels, as well as options for Pub/Sub integration. A 'Cost trend' chart on the right shows spending from February 2024 to 28 February 2025, with a total cost of \$10K.

Google Cloud

Billing / Budgets and alerts / Edit budget

Billing account: Argolis Billing (alexhoang)

Overview

Cost management

- Reports
- Cost table
- Cost breakdown
- Budgets & alerts**
- Billing export
- Anomalies

Cost optimisation

- FinOps hub
- Committed-use discounts...
- CUD analysis
- Pricing
- Cost estimation
- Credits

Payments

- Transactions
- Payment settings

Billing management

- Account management

Amount

Set a monthly budget amount

Budget type: Specified amount

A fixed amount that your spend will be compared against.

Target amount *: \$10000

Actions

Set alert threshold rules

Send email alert notifications after the actual or forecasted spend exceeds a percentage of the budget or a specified amount. [Learn more.](#)

Percentage of budget: 50	Amount 1 *: \$5000	Trigger on 1: Actual
Percentage of budget: 90	Amount 2 *: \$9000	Trigger on 2: Actual
Percentage of budget: 100	Amount 3 *: \$10000	Trigger on 3: Actual

[+ Add threshold](#)

Cost trend

1 February 2024 – 28 February 2025

\$10K

\$5K

Feb Mar Apr May Jun Jul Aug Sept Oct Nov Dec Jan Feb

Total cost

[View report](#)

Manage notifications

Email alerts to billing admins and users

Email alerts to project owners

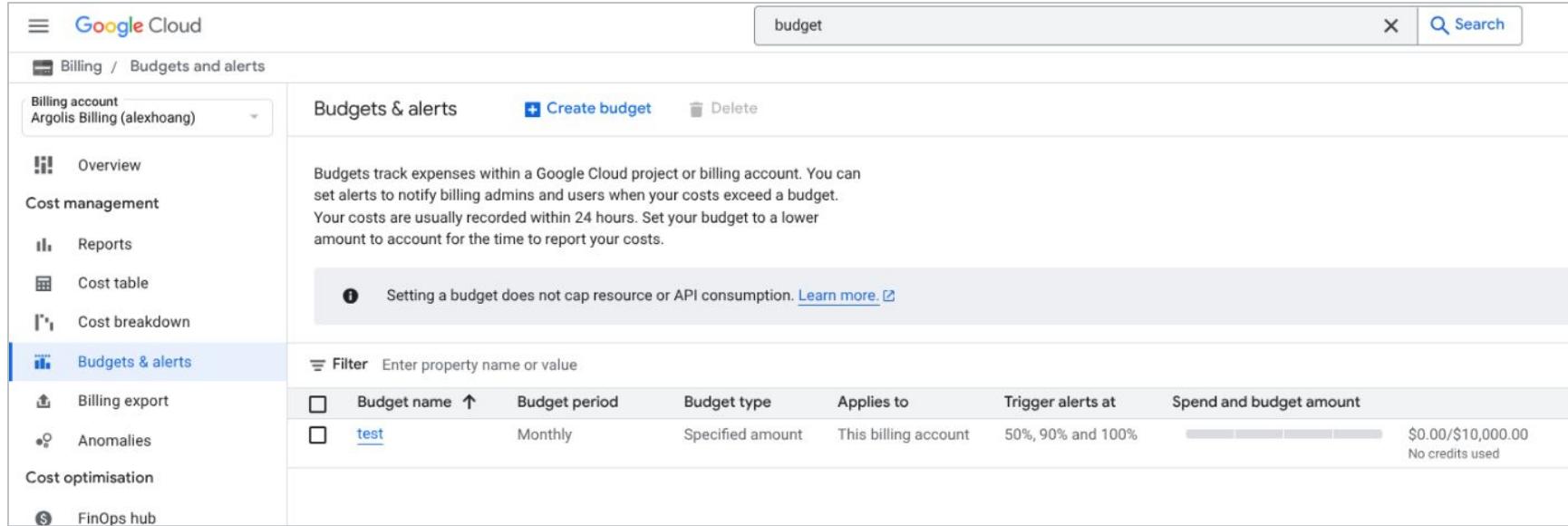
Link Monitoring email notification channels to this budget

Select a project and a maximum of 5 Monitoring email notification channels.

Connect a Pub/Sub topic to this budget

Select a project and Pub/Sub topic. Anyone who can view this budget will also be able to view the project ID and the topic name. It may not be possible to add a Pub/Sub topic if it belongs to an organisation that has [domain-restricted sharing](#) enabled.

Budget Alert | Status



The screenshot shows the Google Cloud Billing & Budgets interface. The left sidebar includes links for Overview, Cost management, Reports, Cost table, Cost breakdown, and the currently selected **Budgets & alerts**. The main content area displays a table of budgets. One budget named "test" is listed, with details: Budget name: "test", Budget period: "Monthly", Budget type: "Specified amount", Applies to: "This billing account", Trigger alerts at: "50%, 90% and 100%", and Spend and budget amount: "\$0.00/\$10,000.00". A note states, "Setting a budget does not cap resource or API consumption." A progress bar indicates "No credits used".

Budget name	Budget period	Budget type	Applies to	Trigger alerts at	Spend and budget amount
test	Monthly	Specified amount	This billing account	50%, 90% and 100%	\$0.00/\$10,000.00 No credits used

- View a list of your budgets and their status.
- Monitor spending with the progress bar.
- Easily modify or delete existing budgets.

You can [set up programmatic notifications](#) to [disable billing usage based on those notifications](#)

Visualize your spending with Looker Studio



Billing | Visualize your costs with Looker Studio



Billing data can be exported to Bigquery and other analytics tools to craft custom dashboards that dive deeper into your cloud usage and costs.

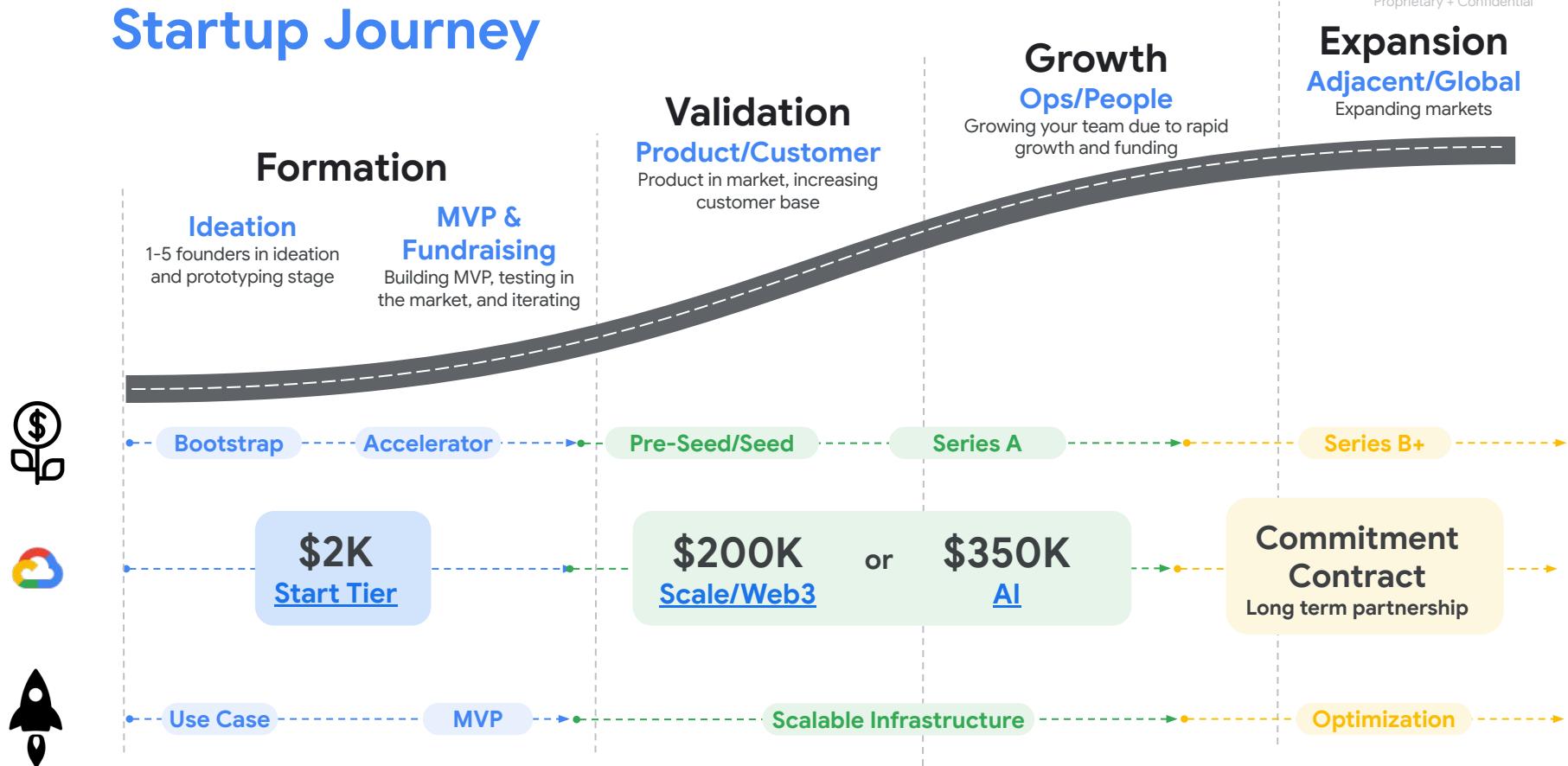
Template: bit.ly/billboard-template



Google for Startups program journey



Startup Journey



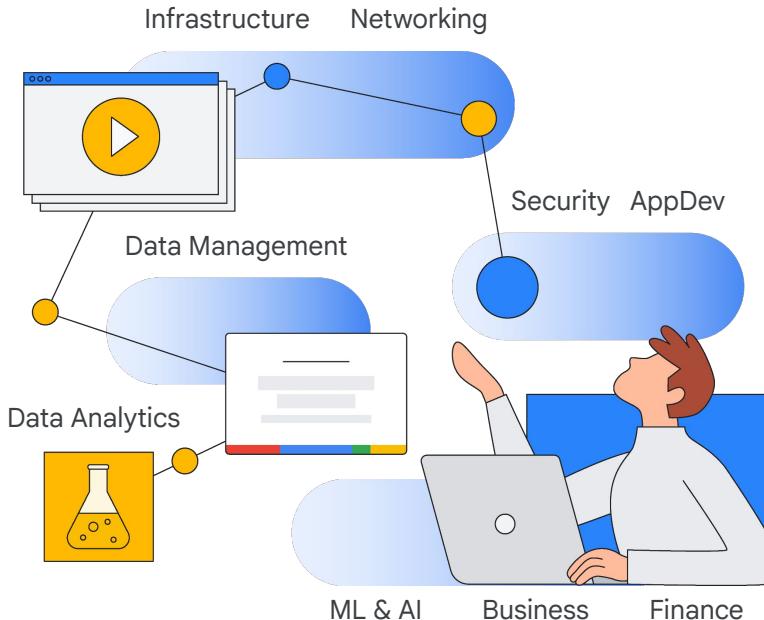
Learn more and apply at cloud.google.com/startup



Request free
on-demand training
licenses

Google Cloud Skills Boost for Organizations

Drive scalable learning for your team with intuitive administrative tools at no cost.



Guided hands-on labs

Build confidence and skills by practicing in a controlled environment

Intuitive administrative tools

Track your team's progress through administrative tools that make it easy to manage your team's learning and key skilling metrics

Skills validation

Measure and recognize cloud capabilities with Google Cloud Credentials

Content by Google Cloud

Content is designed, delivered, and maintained by Google Cloud

No-cost

Up to 500 users at no cost per organization. (Previously \$399/user USD). Additional users available upon request.

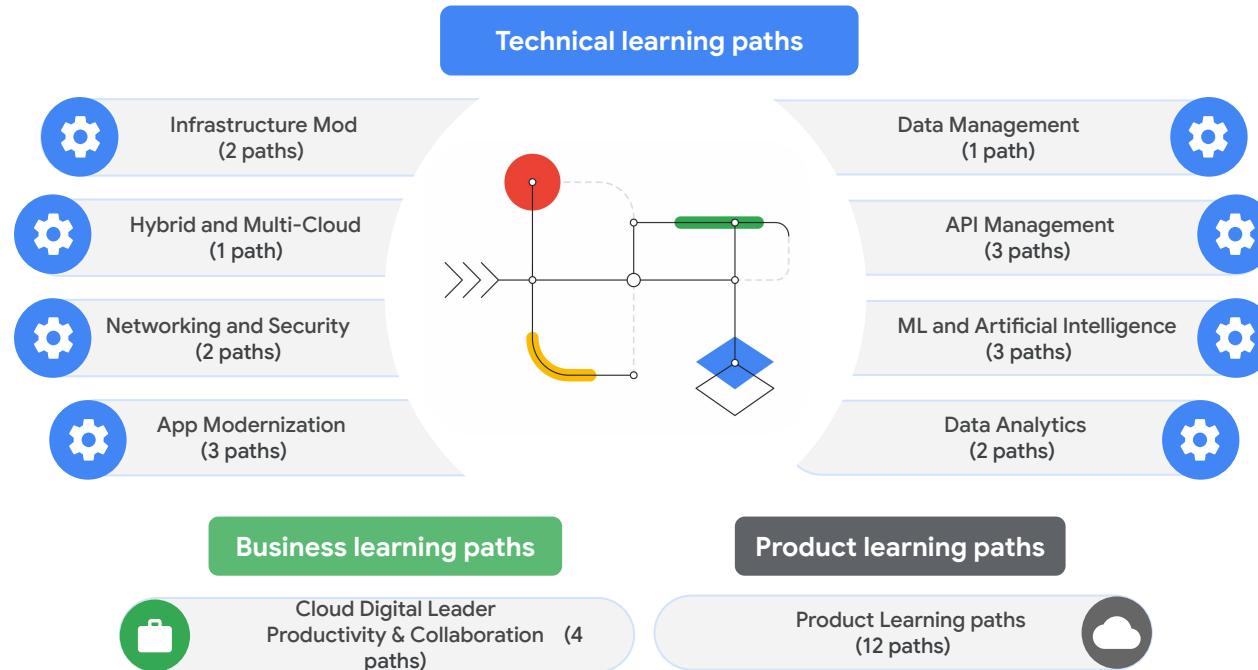
Curated catalog of content

Content created for a range of technical proficiency levels and addresses different learning styles - on-demand videos and courses, self-paced labs, and skill badges

If you are interested, email your account team letting them know the amount of licenses that your team needs.

Learning Paths Cloud Skills Boost

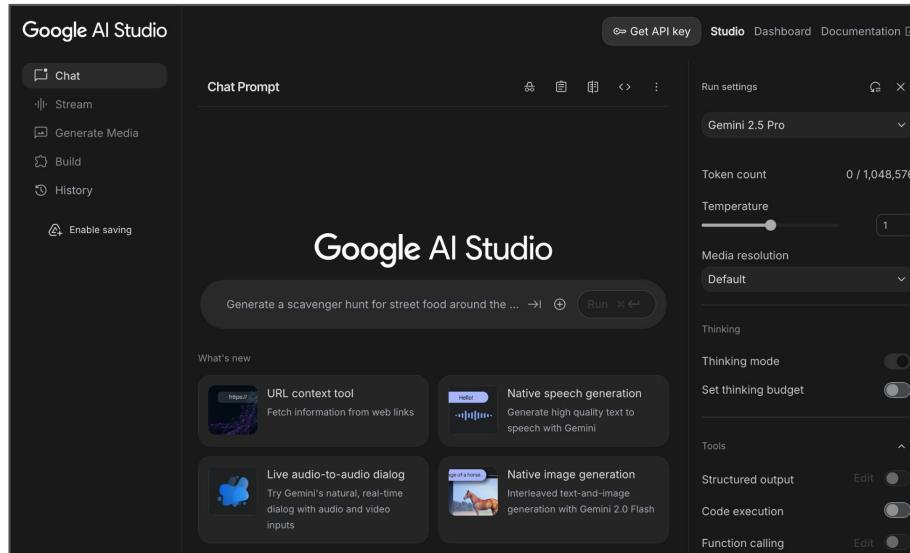
Our [learning paths](#) are curated collections of on-demand content for all business & technical proficiency levels



Access the Gemini APIs



Where can I access the Gemini APIs?



Google AI Studio

Chat Prompt

Run settings

Gemini 2.5 Pro

Token count: 0 / 1,048,576

Temperature: 1

Media resolution: Default

Thinking

Thinking mode: Off

Set thinking budget: Off

Tools

Structured output: Off

Code execution: Off

Function calling: Off

What's new:

- URL context tool: Fetch information from web links
- Native speech generation: Generate high quality text to speech with Gemini
- Live audio-to-audio dialog: Try Gemini's natural, real-time dialog with audio and video inputs
- Native image generation: Interleaved text-and-image generation with Gemini 2.0 Flash

Google AI Studio

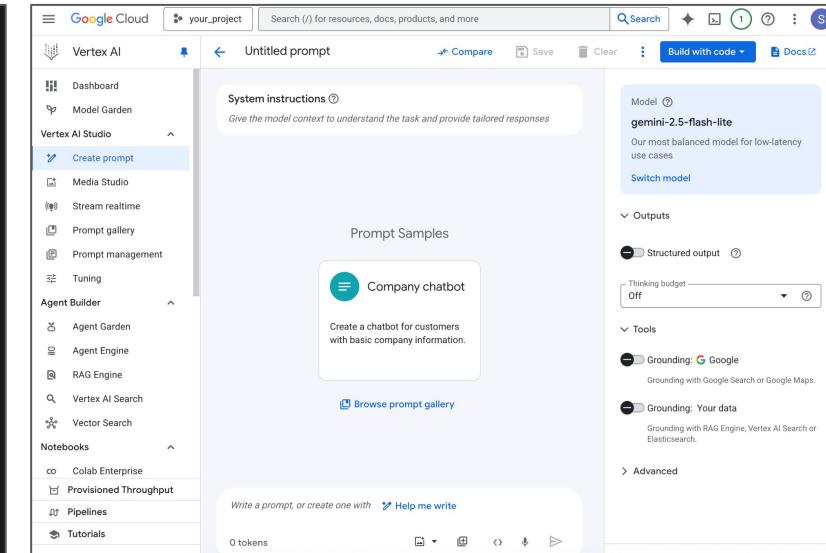
Use case Individual developers and students testing

Quota limits [Spent based rate limits](#)

Multi-region capacity No supported

Reserved capacity No supported

Quickstart [Google AI Studio quickstart](#)



Google Cloud your_project Search (/) for resources, docs, products, and more

Vertex AI

Untitled prompt

System instructions: Give the model context to understand the task and provide tailored responses

Model: gemini-2.5-flash-lite

Switch model

Outputs

Structured output: Off

Thinking budget: Off

Tools

Grounding: Google

Grounding with Google Search or Google Maps.

Grounding: Your data

Grounding with RAG Engine, Vertex AI Search or Elasticsearch.

Prompt Samples

Company chatbot: Create a chatbot for customers with basic company information.

Browse prompt gallery

Write a prompt, or create one with [Help me write](#)

0 tokens

Vertex AI (Recommended)

Production Workloads. Enterprise grade platform.

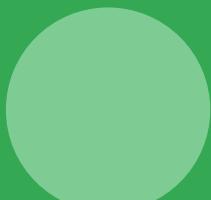
No RPM or TPM limits from day 1 due to [Dynamic Shared Quota](#)

[Vertex AI Global Endpoint](#) leverages multi-region capacity which can help you reduce [429 errors](#) in production

[Provisioned Throughput](#)

[Gemini API in Vertex AI quickstart](#)

Use AI Hypercomputers

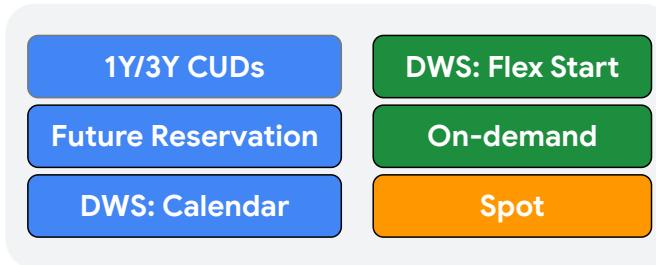


AI Hypercomputers in Google Cloud

Proprietary + Confidential

AI Hypercomputer is a supercomputing system that is optimized to support your artificial intelligence (AI) and machine learning (ML) workloads. AI Hypercomputer uses best practices and systems-level designs to boost efficiency and productivity across AI pre-training, tuning, and serving.

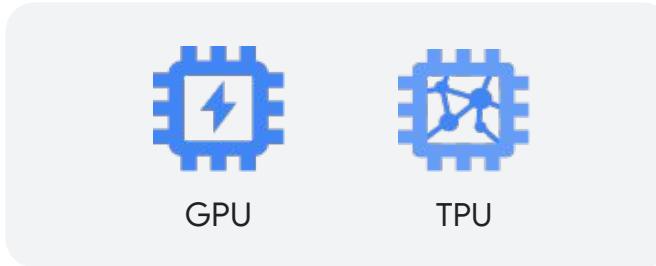
To get started, you need to follow the next steps:



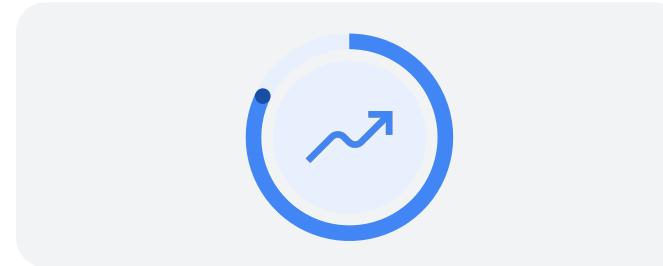
1- Choose your consumption model



2- Select your deployment method



3- Select your accelerator



4- Get your GPU quota limit increased

Google Cloud

Consumption options

Documentation | GitHub | Help

Reserved capacity

Shared capacity

Preemptible

1Y/3Y CUDs

Future Reservation

DWS: Calendar

DWS: Flex Start

On-demand

Spot

Analogy

Renting an apartment

Subletting an apartment

Booking a hotel

Waiting in line

Window shopping

“While supplies last”

Learn more

[Resource-based committed use discounts](#)

[About Future reservations](#)

[About future Reservation in Calendar Mode](#)

[About DWS: Flex-Start](#)

[About Standard provisioning](#)

[Preemptible VM instances](#)

Start Time

Next day 12 AM PT

Future date

Min 24 hours in advance

As soon as available

Very limited

As soon as available

End Time

Fixed. Cannot cancel

Cannot cancel but can be extended

Fixed. Cannot cancel or extend

Cancel any time

Stop any time

Stopped by Google at any time

Duration

1 yr, 3 yr
Can be extended

Min: 24 hours
Max: no limit

Min: 1 day
Max: 90 days

Min: 1 min
Max: 7 days

Min: 1 min
Max: no limit

Min: 1 min
Max: no limit

Pricing

[Discounted](#) (based on duration)

[On demand](#)

[Variable](#) (up to 53% off from On demand)

[Variable](#) (up to 53% off from On demand)

[On demand](#)

[Spot pricing](#)

Deployment

[Purchase commitments with attached reservations](#)

[Create future reservation request](#)

[Create a future reservation request in calendar mode](#)

1- **GCE**: [MIG Resize Request](#)
2- **GKE**: [small batch workload](#) & [large-scale workload](#)
3- **Slurm**: [Cluster Toolkit](#)
4- **Vertex AI**: [Training & Inference](#)
5- **Batch**: [jobs with GPUs](#)

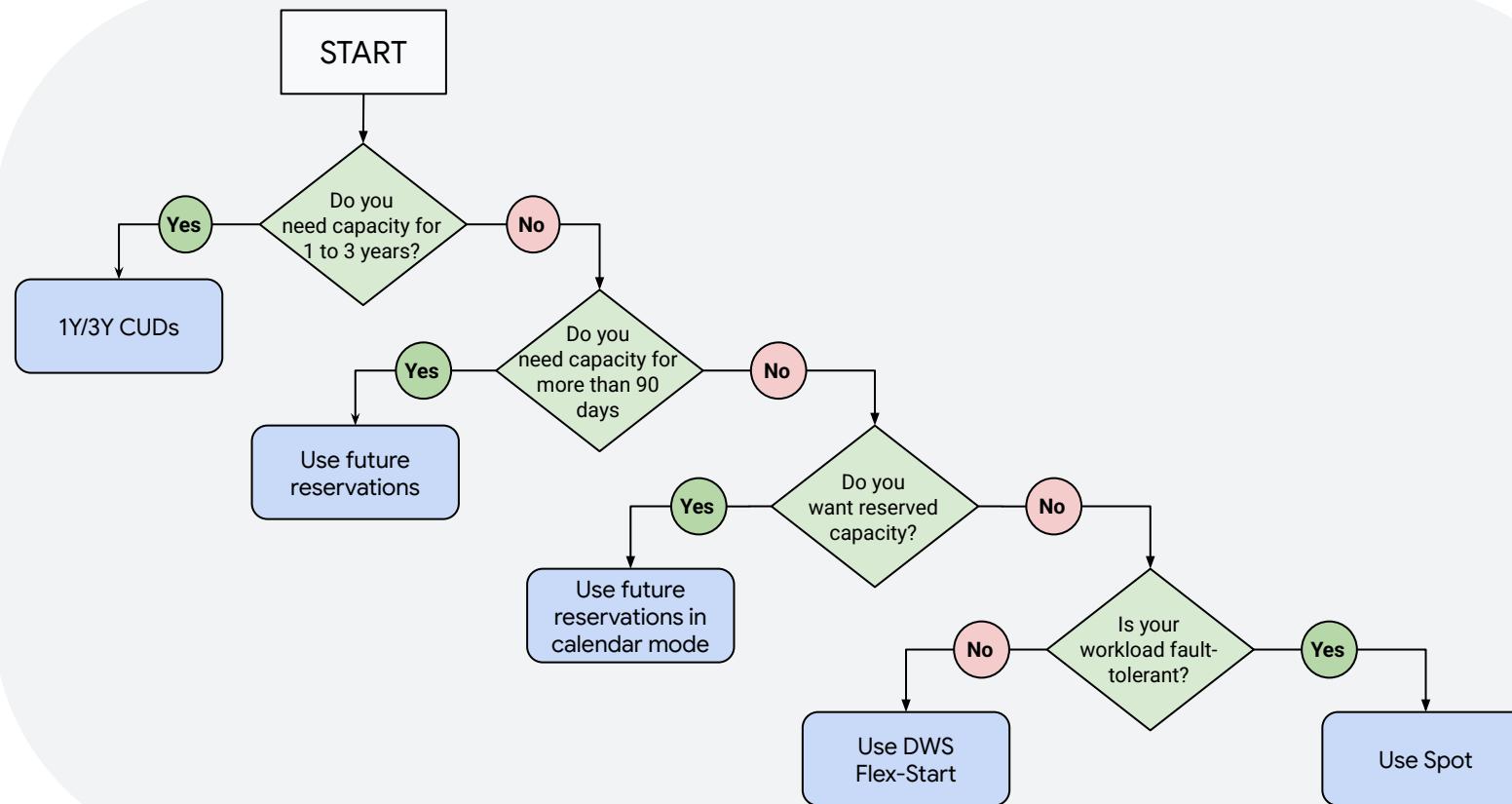
You can request on-demand VMs directly from [Compute Engine](#) or the deployment method of your choice.

You can request spot VMs directly from [Compute Engine](#) or the deployment method of your choice.

Google Cloud

How to choose between the consumption options?

Proprietary + Confidential



Supported Accelerators



Click on the resource names to review their specs



In which regions are those resources supported?

- **Compute Engine**
 - [GPU regions and zones](#)
 - [TPU regions and zones](#)
- **Vertex AI:**
 - [GPU and TPU regions](#)

How to request higher quota limits?

- **Compute Engine:** [Request a quota adjustment](#)
- **Cloud Run:** [Request a quota increase](#)
- **Vertex AI:** Next page →



How to increase GPU quota limits for Vertex AI?

Vertex AI has GPU quota limits separate from Compute Engine.

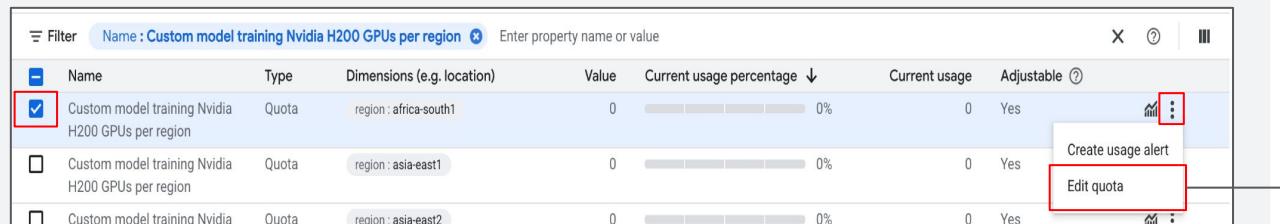
Please follow the instructions below to get GPU quota limits specifically for Vertex AI:

1. Visit [Cloud Console > APIs & Services > Vertex AI > Quotas](#)

2. Use the **Filter** to search for the GPU type you need. Here are some examples for naming conventions:

- Custom model **training** Nvidia H200 GPUs per region [[Vertex AI Training](#)]
- Custom model **serving** Nvidia H200 GPUs per region [[Vertex AI Online and Batch Inference](#)]
- Custom model training **preemptible** Nvidia H200 GPUs per region [[Vertex AI Training + DWS-Flex](#)]
- Custom model serving **preemptible** Nvidia H200 GPUs per region [[Vertex AI Inference + DWS-Flex](#)]

3. Select the **checkbox** next to the quota that you want to update and click **Edit quota**

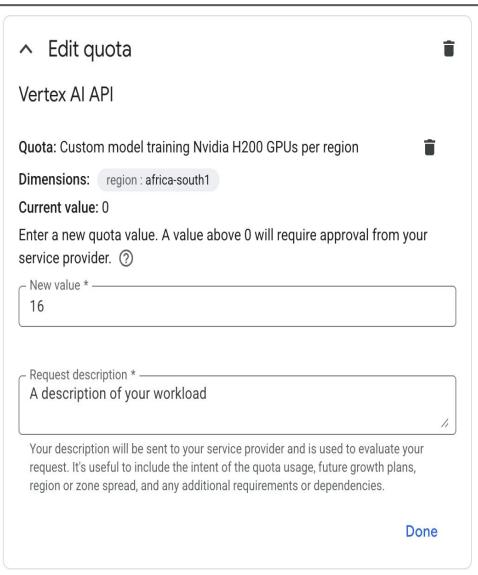


Name	Type	Dimensions (e.g. location)	Value	Current usage percentage	Current usage	Adjustable	⋮
<input checked="" type="checkbox"/> Custom model training Nvidia H200 GPUs per region	Quota	region : africa-south1	0	0%	0	Yes	
<input type="checkbox"/> Custom model training Nvidia H200 GPUs per region	Quota	region : asia-east1	0	0%	0	Yes	
<input type="checkbox"/> Custom model training Nvidia H200 GPUs per region	Quota	region : asia-east2	0	0%	0	Yes	

4. Input the **desired GPU quota limit** and provide a workload description for us to review. Click **Next**.

5. Provide us a **contact information** to keep you updated on the request progress. **Submit request**.

6. **To expedite your request**, let your Account Team know that you requested GPUs for Vertex AI and send them the project-ids to locate your requests.



^ Edit quota

Vertex AI API

Quota: Custom model training Nvidia H200 GPUs per region

Dimensions: region : africa-south1

Current value: 0

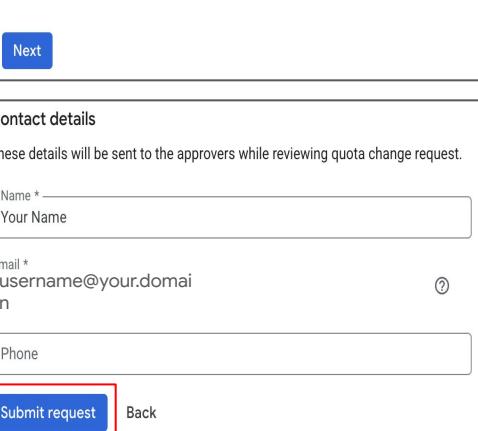
Enter a new quota value. A value above 0 will require approval from your service provider. ?

New value *

Request description *
A description of your workload

Your description will be sent to your service provider and is used to evaluate your request. It's useful to include the intent of the quota usage, future growth plans, region or zone spread, and any additional requirements or dependencies.

Done



Next

Contact details

These details will be sent to the approvers while reviewing quota change request.

Name *

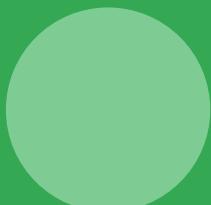
Email *

Phone

Submit request

Back

File a support case



Support | Navigate to the support cases page

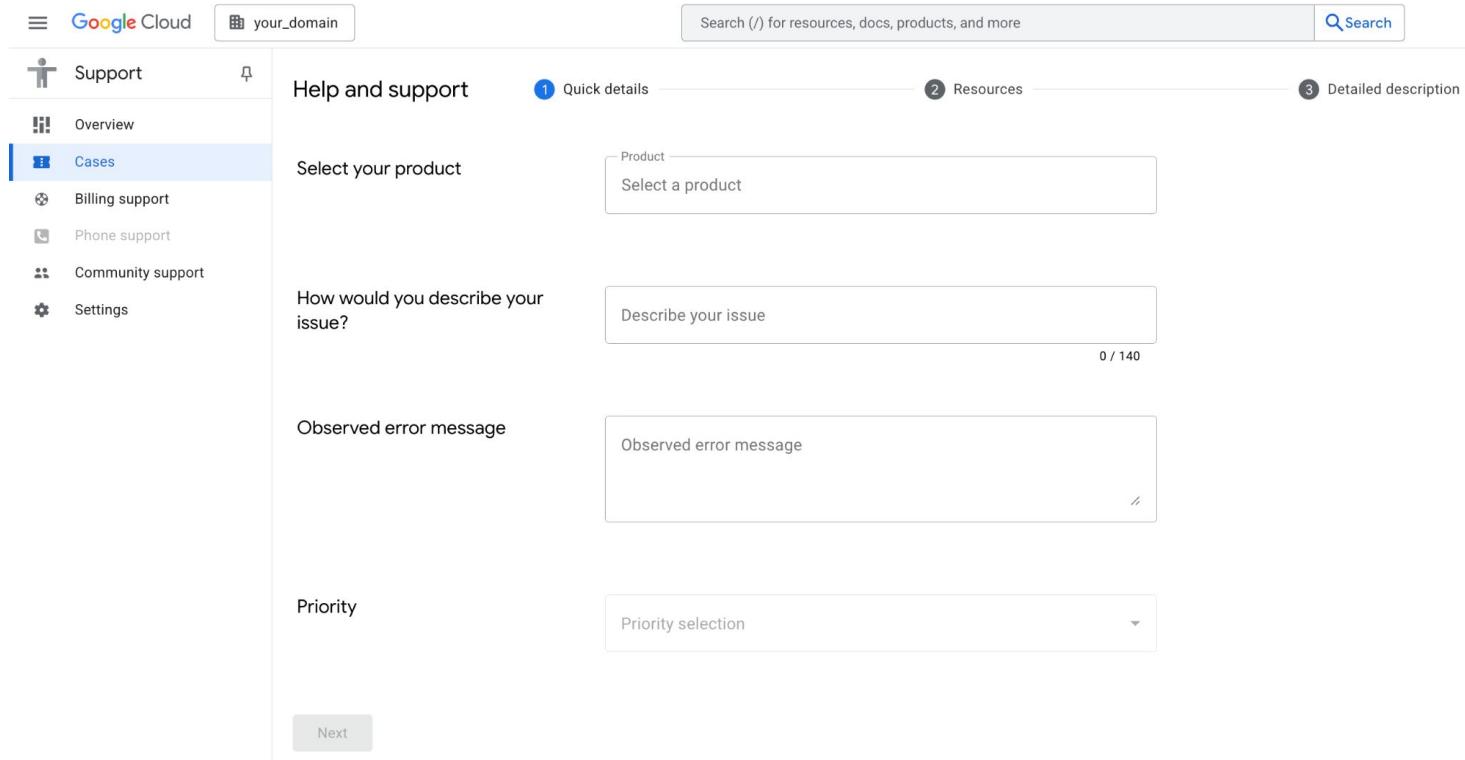
console.cloud.google.com/support/cases

If "GET HELP" is grayed out, review your IAM roles again.

Support cases will be listed here.

Title	Priority	Status	Project	Case Number	Updated
-------	----------	--------	---------	-------------	---------

Support | Fill in the form and submit it



The screenshot shows the Google Cloud Support interface for creating a new case. The left sidebar has a 'Cases' link selected. The main area is a 'Help and support' form with three tabs: 'Quick details' (selected), 'Resources', and 'Detailed description'. The 'Quick details' tab contains fields for 'Select your product' (a dropdown menu with 'Select a product'), 'How would you describe your issue?' (a text area with 'Describe your issue' placeholder and character count '0 / 140'), 'Observed error message' (a text area with 'Observed error message' placeholder and character count '0 / 140'), and 'Priority' (a dropdown menu with 'Priority selection' placeholder). A 'Next' button is at the bottom.

Google Cloud your_domain

Search (/) for resources, docs, products, and more

Support

Help and support

① Quick details

② Resources

③ Detailed description

Select your product

Product

Select a product

How would you describe your issue?

Describe your issue

0 / 140

Observed error message

Observed error message

Priority

Priority selection

Next

The following table defines support case priorities. For more information, see [Best practices for working with Support](#)

Priority definition	Example situations
P1: Critical Impact—Service Unusable in Production	<p>The application or infrastructure is unusable in production, having a significant rate of user-facing errors.</p> <p>Business impact is critical (for example, revenue loss or potential data integrity issue).</p> <p>No workaround is available that can be quickly implemented (less than 30 minutes).</p> <p>Affected Google Cloud component or feature is marked as General Availability.</p> <p>Immediate attention from Google is required to resolve the problem.</p>
P2: High Impact—Service Use Severely Impaired	<p>The infrastructure is degraded in production, having a noticeable rate of user-facing errors or difficulties in spinning up a new production system.</p> <p>Business impact is moderate (for example, danger of revenue loss or productivity decrease).</p> <p>A workaround to mitigate critical business impact is available and can be quickly implemented.</p> <p>Affected Google Cloud component or feature is marked as General Availability.</p> <p>Fast response from Google is required.</p>
P3: Medium Impact—Service Use Partially Impaired	<p>The issue is limited in scope and/or severity. The issue has no user-visible impact.</p> <p>Business impact is low (for example, inconvenience or minor business processes affected).</p> <p>Case requires more in-depth investigation and troubleshooting and less frequent interactions.</p>
P4: Low Impact—Service Fully Usable	<p>Little to no business or technical impact.</p> <p>Recommended for consultative tickets where in-depth analysis, troubleshooting or consultancy are preferred to more frequent communications.</p>

Next steps



Google Cloud

Next steps

- 01. Let your **Customer Engineer** know about any capacity needs, blockers, or technical questions you may have.
- 02. Connect with your **FSR** for any questions about the Google for Startups program.
- 03. Keep your **account team** (FSR/CE) in the loop for any launch or milestone updates.

Thank you

Join us as we keep pushing boundaries at
Next 26 on April 22–24, 2026
at Mandalay Bay Convention Center in Las Vegas
cloud.withgoogle.com/next

