Module 6

Scaling with Google Cloud Operations

Lessons

01 Financial governance and managing cloud costs

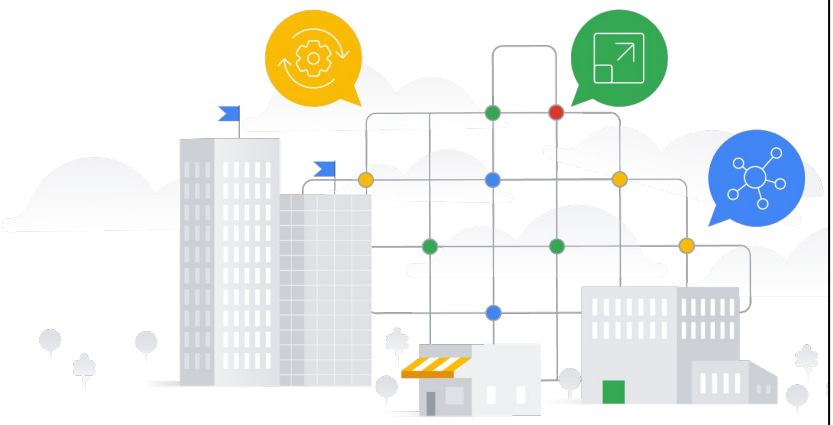
02 Operational excellence and reliability at scale

03 Sustainability with Google Cloud

Say: Welcome to module 6, the final section of this Cloud Digital Leader training.

Managing and scaling cloud resources effectively can be a complex task

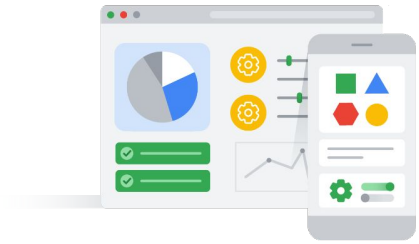
Cloud operations refers to the set of practices and strategies employed to ensure the smooth functioning, optimization, and scalability of cloud-based systems.



Say: In today's digital landscape, organizations of all sizes are embracing the power and flexibility of the cloud to transform how they operate. However, managing and scaling cloud resources effectively can be a complex task. That's where cloud operations come in.

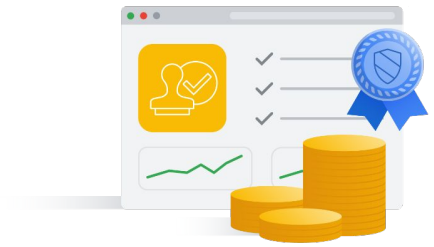
Cloud operations refers to the set of practices and strategies employed to ensure the smooth functioning, optimization, and scalability of cloud-based systems.

Cloud operations



Managing and monitoring:

- Infrastructure
- Applications
- Services



Adhering to best practices:

- Reliability
- Performance
- Security
- Cost optimization

Say: It involves managing and monitoring the infrastructure, applications, and services that run in the cloud, while adhering to best practices for reliability, performance, security, and cost optimization.

Cloud operations play a pivotal role in enabling organizations to achieve digital transformation goals, because they ensure the availability, efficiency, and resilience of critical systems.

So, with this in mind, let's explore the goals of the **final** section of this course. "Scaling with Google Cloud Operations" was designed to help you:

- Learn how Google Cloud supports an organization's ability to control their cloud costs through financial governance.
- Understand the fundamental concepts of modern operations, reliability, and resilience in the cloud.
- And explore how Google Cloud works to reduce our environmental impact and help organizations meet sustainability goals.

Module 6

Scaling with Google Cloud Operations

Lessons

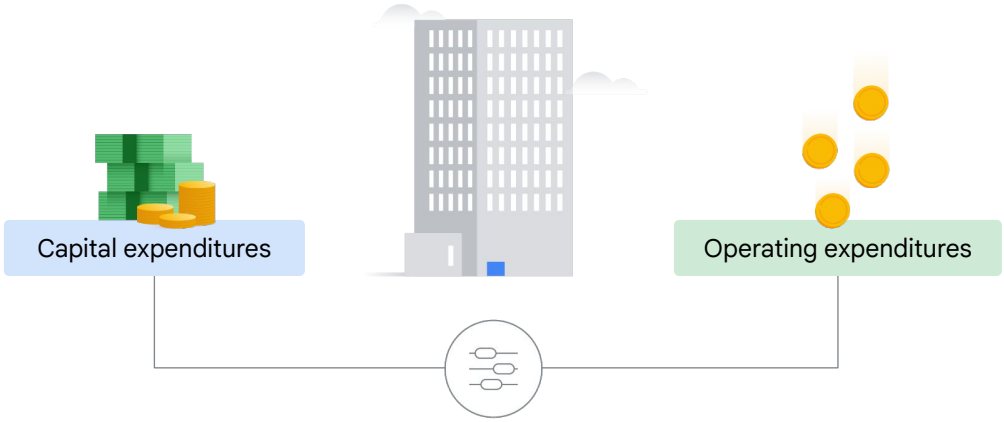
01 Financial governance and managing cloud costs

02 Operational excellence and reliability at scale

03 Sustainability with Google Cloud

Say: Using cloud technology, either for business improvements or for large-scale transformation, can be challenging. In fact, one of the common pain points many organizations face, regardless of which cloud provider they use, is **managing cloud costs**.

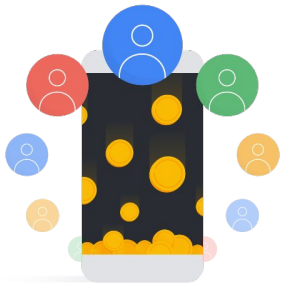
The transition CapEx to OpEx requires process and organizational changes



Say: For large organizations especially, the transition from predictable capital expenditures for building and maintaining their IT infrastructure to agile operating expenditures using cloud resources requires process and organizational changes.

Managing cloud costs also requires vigilance and real-time monitoring

Managing IT infrastructure costs no longer sits mainly with the finance team



It involves more people across multiple teams

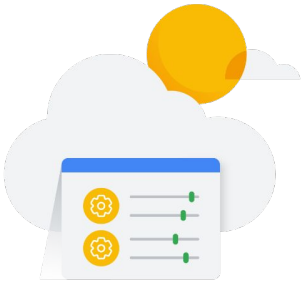
Say: Managing cloud costs requires vigilance and real-time monitoring in parallel. In fact, because almost anyone can now access cloud resources on demand, managing IT infrastructure costs no longer sits mainly with the finance team.

Instead, it involves more people across multiple teams. So you might even be the person responsible for IT budgeting. Whatever your role, understanding how using cloud technology affects the business from a cost perspective will help you maximize the value your organization gains from using the cloud.

Fundamentals of cloud financial governance

Say: With that in mind, let's explore some of the fundamental of cloud financial governance.

Cloud financial governance can mean the difference between peace of mind and spiraling costs



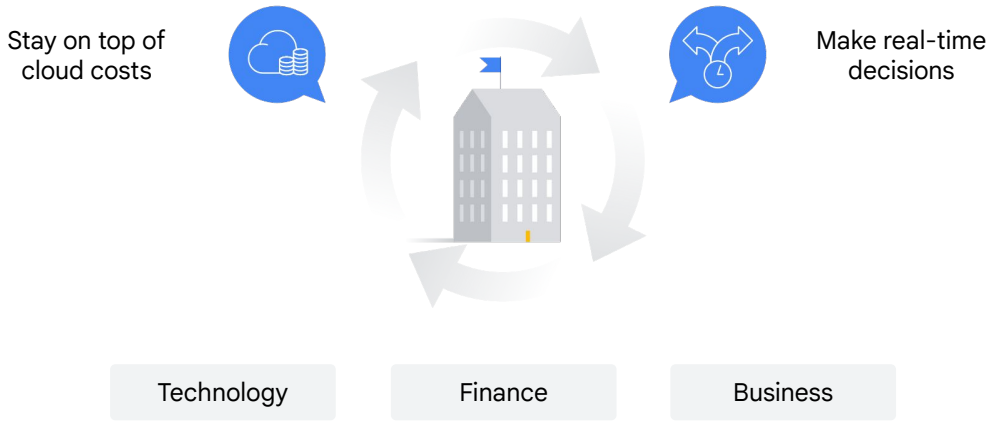
Precise, real-time control of what's being consumed

Cloud financial governance

Say: Easy access to cloud resources presents a need for precise, real-time control of what's being consumed.

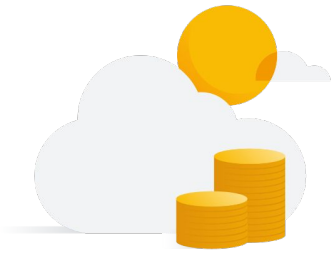
Having cloud financial governance, which is in part a set of processes and controls that organizations use to manage cloud spend, can mean the difference between peace of mind and spiraling costs that lead to budget overruns.

Organizations will need a core team across functions



Say: As an organization adapts, it'll need a core team across technology, finance, and business functions to work together to stay on top of cloud costs and make decisions in real time.

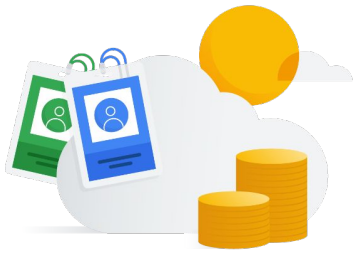
Cloud costs impacts people, process, and technology



People | Process | Technology

Say: The variable nature of cloud costs impacts **people**, **process**, and **technology**. Let's explore these three areas, starting with **people.**

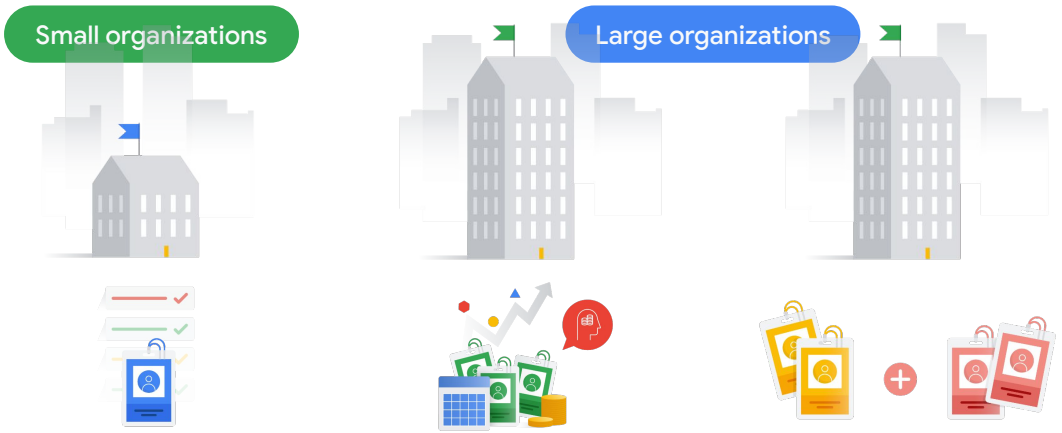
People refers to the different roles involved in managing cloud costs



People | Process | Technology

Say: *People* refers to the different roles involved in managing cloud costs.

The roles people play at organizations of different sizes

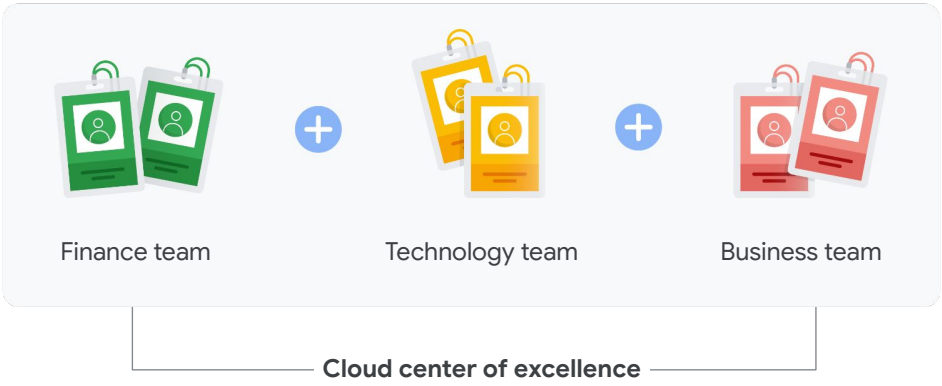


Say: For **small organizations**, one person might fulfill multiple roles and be responsible for managing all aspects of a cloud infrastructure and associated finance. From budgeting to procurement, tracking optimization, and more.

Large organizations, however, will likely look to a finance team to take on a financial planning and advisory role. Using business priorities, a finance team is expected to make data-driven decisions on cloud spending, but they might struggle to understand or monitor cloud spend on a daily, weekly, or monthly basis.

Then there are **members of technology and line of business teams**. They can advise on *how* cloud resources are being used to meet the organization's overall business strategy and what additional resources might be needed throughout the upcoming year. However, they don't necessarily factor costs into their decision making.

To manage cloud costs effectively, a partnership is required



Say: To manage cloud costs effectively, a partnership across finance, technology, and business functions is required. This partnership might already exist, or it may take the form of a centralized hub, such as a cloud center of excellence.

01 Ensure that best practices are in place across the organization.

02 Ensure there's visibility into the ongoing cloud spend.

03 Make real-time decisions and discuss trade-offs.

Cloud center of excellence

Say: The central team would consist of several experts who ensure that best practices are in place across the organization and that there's visibility into the ongoing cloud spend. The centralized group would also be able to make real-time decisions and discuss trade-offs when spending is higher than planned.

Process

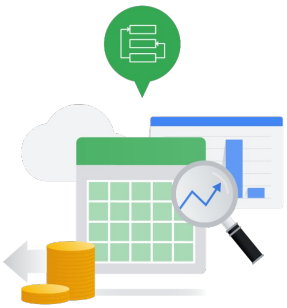
People | **Process** | Technology

Say: Now let's transition from people to **process**.

Processes for organizations to implement

On a daily or weekly basis

Monitor and analyze cloud usage and costs.



On a weekly or monthly basis

Analyze the results and charge back the costs through the appropriate teams. Make changes, if needed.

Say: On a daily or weekly basis, organizations should monitor and analyze their cloud usage and costs.

Then, on a weekly or monthly basis, the finance team should analyze the results, charge back the costs through the appropriate teams, and determine whether any changes are needed to ensure that the organization's cloud spend is optimized.

The importance of having a culture of accountability

- Recognize waste, quickly act to eliminate it, and ensure they're maximizing their cloud investment.
- Drive cross-group collaboration across technology, finance, and business teams to ensure their cloud spend aligns with broader business objectives.

Say: Having a culture of accountability in place across teams helps organizations recognize waste, quickly act to eliminate it, and ensure they're maximizing their cloud investment.

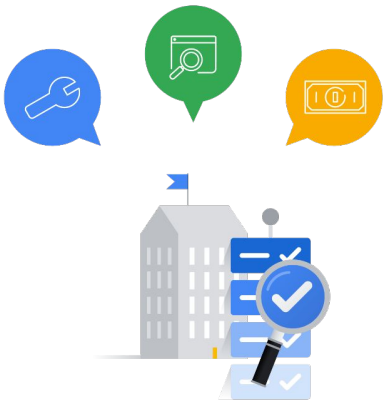
It will also help drive cross-group collaboration across technology, finance, and business teams to ensure that their cloud spend aligns with broader business objectives.

Technology

People Process **Technology**

Say: And finally, there's **technology.**

Google Cloud provides built-in tools to help organizations monitor and manage costs



These tools help organizations:

- Gain greater visibility
- Drive a culture of accountability for cloud spending across the organization
- Control costs to reduce risks of overspending
- Provide intelligent recommendations to optimize cost and usage

Say: Google Cloud provides built-in tools to help organizations monitor and manage costs.

These tools help organizations gain greater visibility, drive a culture of accountability for cloud spending across the organization, control costs to reduce risks of overspending, and provide intelligent recommendations to optimize cost and usage. We'll explore some of these tools later in this section.

Discussion

Financial governance

- What challenges do you face when managing cloud spend?
- What processes or framework does your organization have in place to manage cloud costs?

Say: Let's pause for a quick discussion around financial governance. We're going to explore this in more detail shortly, but first…

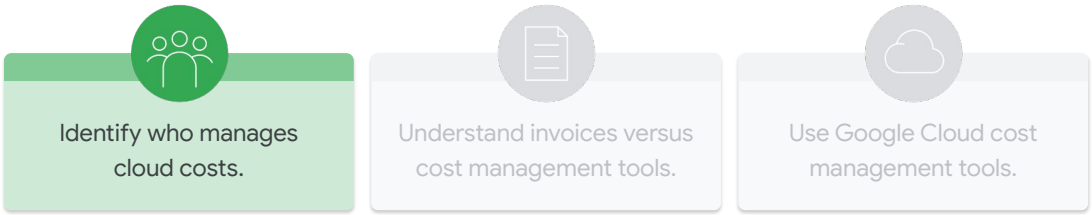
Ask:

- What challenges do you face when managing cloud spend?
- What structure does your organization have in place to manage cloud costs?

Cloud financial governance best practices

Say: Let's explore some cloud financial governance best practices that organizations can adopt to increase the predictability and control of their cloud resources.

Cloud financial governance best practices



Say: The first best practice is to **identify who manages cloud costs**. If it's a team, it should ideally be a mix of IT managers and financial controllers.

Because Cloud spending is decentralized and variable, it's important to establish a culture of accountability for costs across the organization.

Defining clear ownership for projects and sharing cost views with the departments and teams that are using cloud resources helps establish this accountability culture and more responsible spending.

Establish a culture of accountability for costs across the organization

Cloud spending is decentralized and variable.

Define clear ownership for projects.

Share cost views with the departments and teams that are using cloud resources.

Say: Because Cloud spending is decentralized and variable, it's important to establish a culture of accountability for costs across the organization.

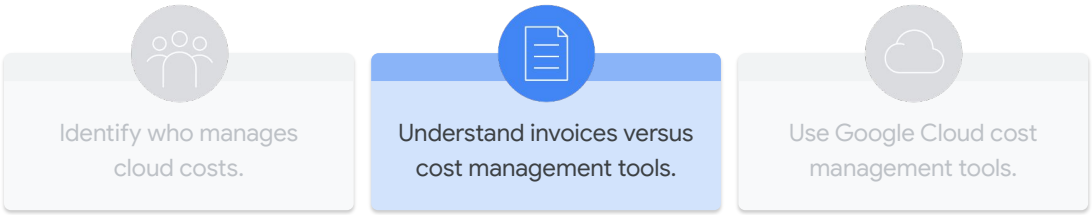
Defining clear ownership for projects and sharing cost views with the departments and teams that are using cloud resources helps establish this accountability culture and more responsible spending.

Policies and permissions help control who can spend and view costs across

Say: As well as making teams accountable for their spending, Google Cloud financial governance policies and permissions make it easy to control who can spend and view costs across your organization. In addition, Google Cloud offers flexible options to organize resources and allocate costs to individual departments and teams.

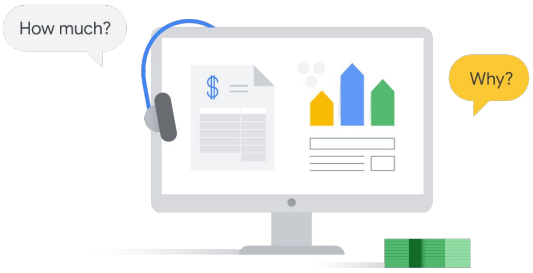
For example, *budgets* notify key stakeholders based on your actual or forecasted cloud costs. Creating multiple budgets with meaningful alerts is an important practice for staying on top of your cloud costs.

Cloud financial governance best practices



Say: The second best practice is to **understand what kind of information can be found in an invoice versus cost management tools**. They're not the same concept.

Invoices vs. cost management tools



Invoice

A document sent by a cloud service provider to a customer to request payment for the services that were used.

Cost management tool

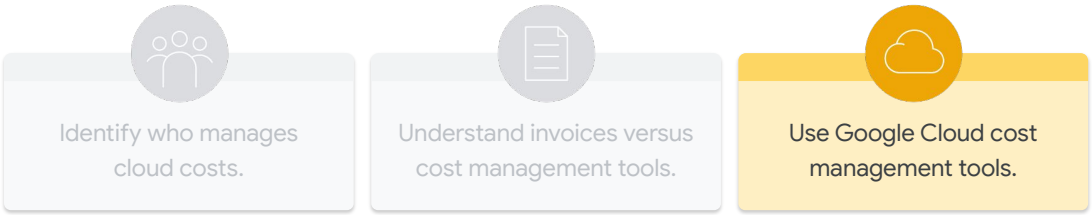
Software to help track, analyze, and optimize cloud spend.

Say: An **invoice** is a document that is sent by a cloud service provider to a customer to request payment for the services that were used. However, a **cost management tool** is software to help track, analyze, and optimize cloud spend.

An organization is rarely *only* interested in how much they spend. They also want to know *why* they spent that much.

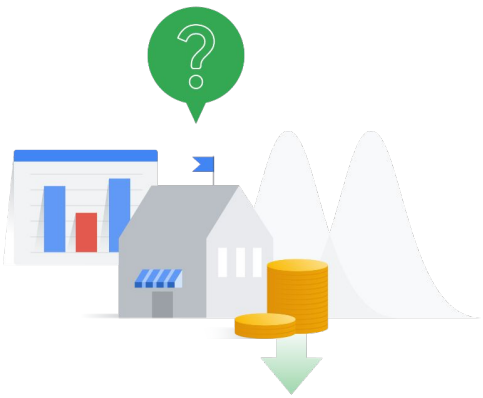
Cost management tools, like those built into the Google Cloud console, are effective for answering the why. They can provide granular data, uncover trends, and identify actions to take to control or optimize costs.

Cloud financial governance best practices



Say: And this brings us to the third best practice for increasing the predictability and control of cloud resources: **use Google Cloud cost management tools.**

Organizations must understand their cloud spend



Capture what cloud resources are being used, by whom, for what purpose, and at what cost.

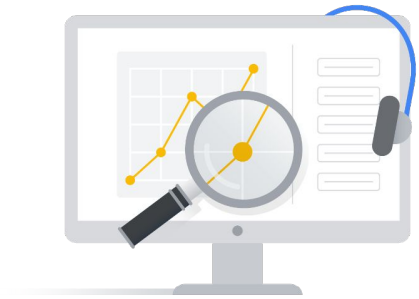
Determine who is responsible for monitoring that information, who is involved in managing costs, and how the spending information is reported on an ongoing basis.

Set up the cadence and format for ongoing communication with main cloud stakeholders.

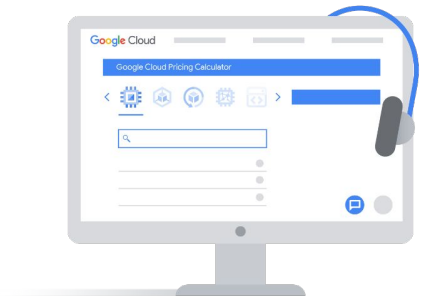
Say: Google Cloud believes in supporting organizations by providing strong financial governance tools that make it easier for customers to align their strategic priorities with their cloud usage. Before organizations can optimize their cloud costs, they first need to understand what they're spending, whether there are any trends, and what their forecasted costs are. So, how can this be done?

- Start by capturing what cloud resources are being used, by whom, for what purpose, and at what cost.
- From there, determine who will be responsible for monitoring that information, who will be involved in managing costs, and how the spending information will be reported on an ongoing basis.
- It's also important to set up the cadence and format for ongoing communication with main cloud stakeholders. Having this plan outlined up front helps ensure that managing costs isn't an afterthought.

Google Cloud helps organizations monitor cost trends and identify areas of waste



Built-in reporting capabilities



Google Cloud Pricing Calculator

cloud.google.com/products/calculator

Say: And how can you monitor current cost trends and identify areas of waste that could be improved?

Google Cloud provides built-in reporting capabilities, which can help your team gain visibility into costs. Ideally, reports should be reviewed weekly, at a minimum.

One powerful tool is the Google Cloud Pricing Calculator. The Pricing Calculator lets you estimate how changes to cloud usage will affect costs. The calculator is available at **cloud.google.com/products/calculator**.

Activity

10 min | Class

Go to cloud.google.com/products/calculator

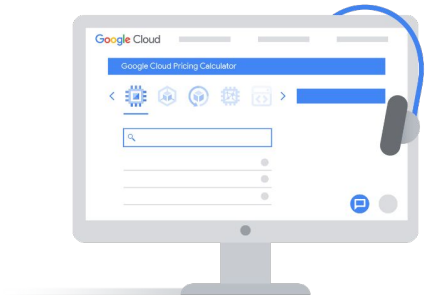
On the slides that follow, you'll need to add the listed products to the Pricing Calculator and see if your final total in USD matches those of others in the class.

Don't change any product configurations unless specified.

Say: Now let's try out the Pricing Calculator. For the purposes of this activity, we are going to practice navigating the calculator tool. You'll likely see lots of new terms, and we could probably spend an entire day talking through different configurations, but that's not our focus right now.

Read: Go to cloud.google.com/products/calculator. On the slides that follow, you'll need to add listed products to the Pricing Calculator and see if your final total in USD matches those of others in the class. Don't change any product configurations unless specified.

Scenario 1



Google Cloud Pricing Calculator

Add:

2x Compute Engine instances with sustained use discounts in the Las Vegas (us-west4) region.

Say: Add 2x Compute Engine instances with sustained use discounts in the Las Vegas (us-west4) region.

Scenario 1 - Compute

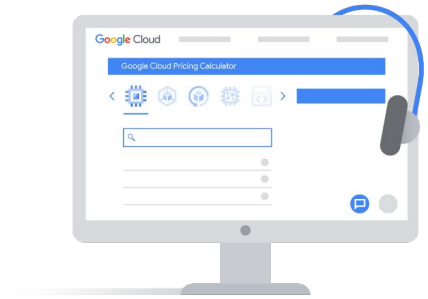
Instances (Compute Engine)		$220.45
Service type	Instances	
Instance-time	1460 Hours	N/A
Machine type	n1-standard-4, vCPUs: 4, RAM: 15 GB	$218.69
Boot disk type	Standard persistent disk	N/A
Boot disk size (GiB)	20 GiB	$1.76
Number of Instances	2	N/A
Operating System / Software	Free: Debian, CentOS, CoreOS, Ubuntu or BYOL (Bring Your Own License)	N/A
Provisioning Model	Regular	N/A
Threads per core	2 threads per core	N/A
Enable Confidential VM service	false	N/A
Add sustained use discounts	true	N/A
Add GPUs	false	N/A
Local SSD	0	N/A
Region	Las Vegas (us-west4)	N/A
Committed use discount options	None	N/A

Say: Compute summary shows that 2 Compute Engine instances have been selected, along with sustained use discounts.

In this example the price for this product and configuration comes to **$220.45**.

Ask: What total did you get for the compute part of the exercise?

Scenario 2



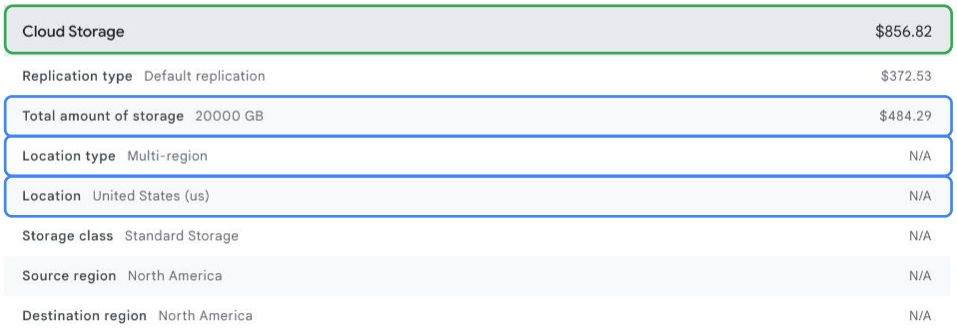
Google Cloud Pricing Calculator

Add:

20,000 GB of multi-region Cloud Storage in the US

Say: Add 20,000 GB of multi-region Cloud Storage in the US.

Scenario 2 - Storage

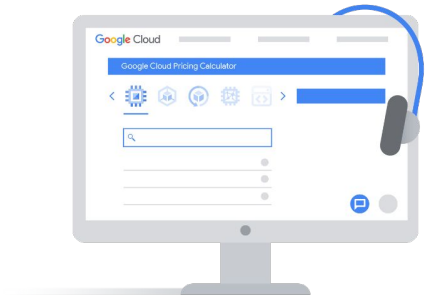


Say: Storage summary shows 20000 GB of storage has been selected, located in the United States multi-region.

In this example the price for this product and configuration comes to **$856.82**.

Ask: What total did you get for the storage part of the exercise?

Scenario 3



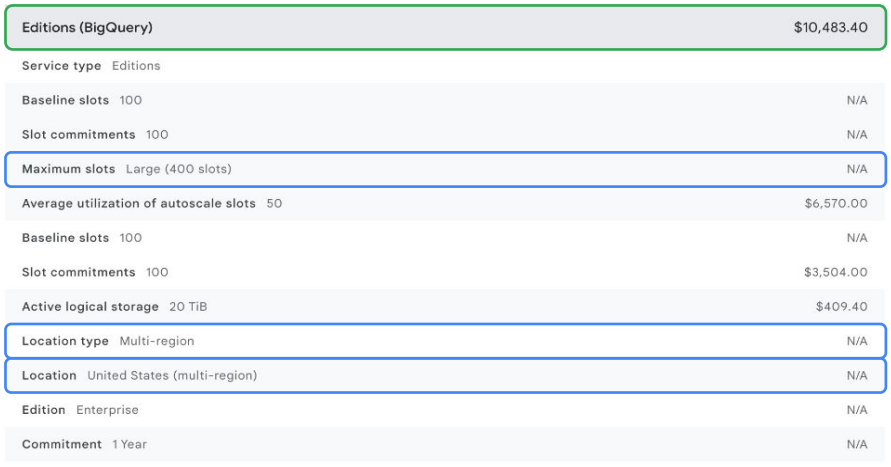
Google Cloud Pricing Calculator

Add:

A multi-region BigQuery instance in the US with 400 slots

Say: Say: Add a multi-region BigQuery instance in the US with 400 slots.

Scenario 3 - Data analytics



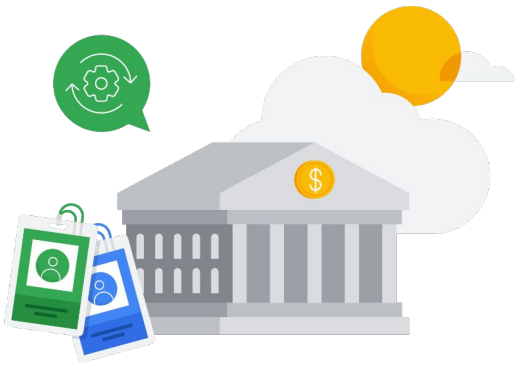
Editions (BigQuery)	$10,483.40
Service type Editions	
Baseline slots 100	N/A
Slot commitments 100	N/A
Maximum slots Large (400 slots)	N/A
Average utilization of autoscale slots 50	$6,570.00
Baseline slots 100	N/A
Slot commitments 100	$3,504.00
Active logical storage 20 TiB	$409.40
Location type Multi-region	N/A
Location United States (multi-region)	N/A
Edition Enterprise	N/A
Commitment 1 Year	N/A

Say: Data analytics summary shows that 400 slots have been configured in the United States multi-region.

In this example the price for this product and configuration comes to **$10,483.40**.

Ask: What total did you get for the data analytics part of the exercise? Also, what's your **total price** and how does it compare to others in the class?

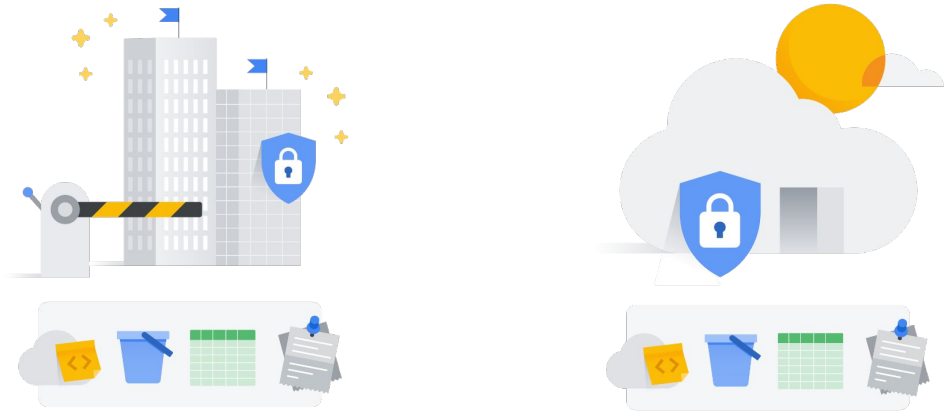
The next step: Implement the financial governance best practices



Say: Now that you've had a chance to explore some cloud financial governance best practices, the next step is to implement them. If this doesn't fall into your scope of responsibility, be sure to pass on those best practices to the relevant stakeholders within your organization.

Using the resource hierarchy to control access

Controlling access to resources in the cloud requires different methods



Say: One important cloud computing consideration involves controlling access to resources.

With on-premises infrastructure, physical access controls were used. This method, however, is not as effective with resources stored in the cloud.

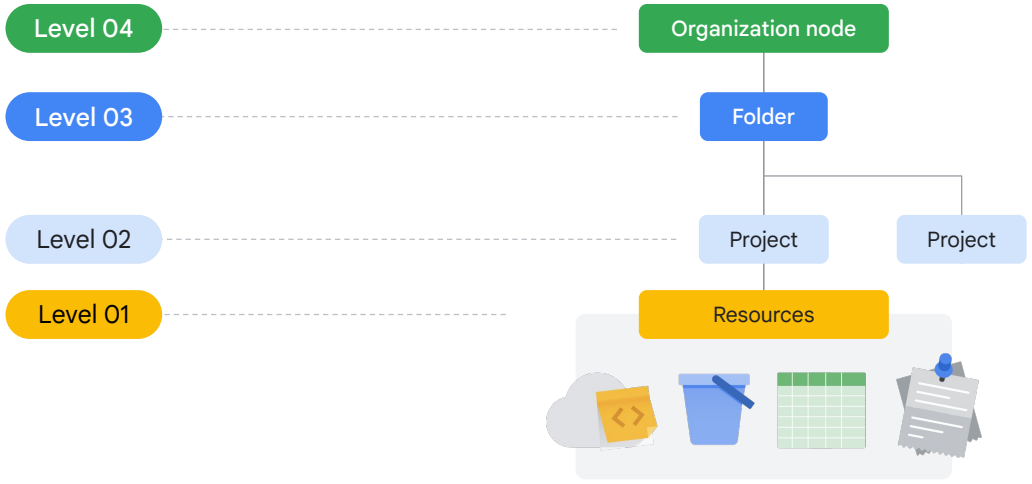
The Google Cloud resource hierarchy

- It's a powerful tool that can be used to control access to cloud resources.
- This tree-like structure organizes resources into logical groups.

Say: The Google Cloud resource hierarchy is a powerful tool that can be used to control access to cloud resources.

Much like the folder structure you use to organize and control access to your own files, this resource hierarchy is a tree-like structure that organizes resources into logical groups. This makes it easier to manage resources and control.

The Google Cloud resource hierarchy



Say: Google Cloud's resource hierarchy contains four levels, and starting from the bottom up they are: resources, projects, folders, and an organization node.

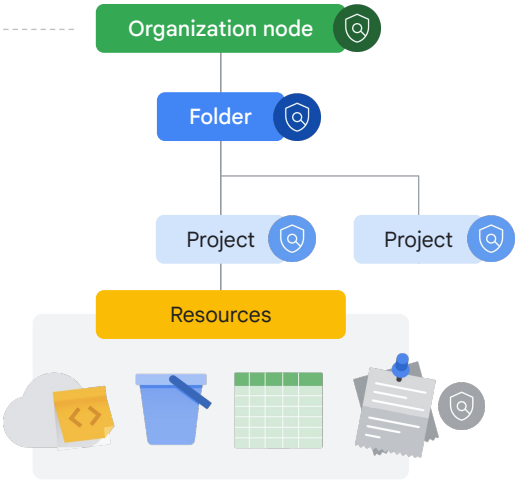
1. The first level, **resources**, represent virtual machines, Cloud Storage buckets, tables in BigQuery, or anything else in Google Cloud.
2. Resources are organized into **projects**, which sit on the second level.
3. Projects can be organized into **folders**, or even subfolders. These sit at the third level.
4. And then at the top level is an **organization node**, which encompasses all the projects, folders, and resources in your organization.

It's important to understand this resource hierarchy because it directly relates to how policies are managed and applied when you use Google Cloud.

Policies

Add Policy

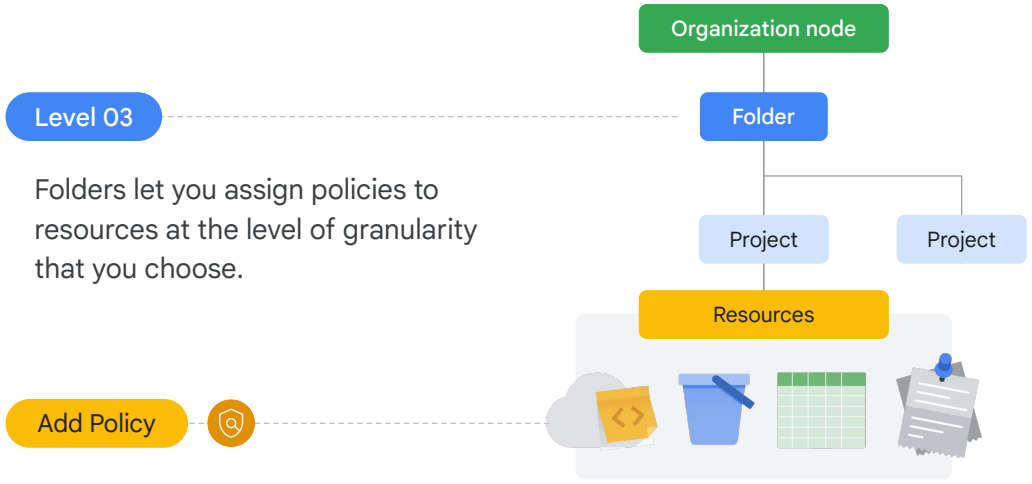
A set of rules that define **who** can access a resource and **what** they can do with it.



Say: A policy is a set of rules that define *who* can access a resource and *what* they can do with it.

Policies can be defined at the project, folder, and organization node levels. Some Google Cloud services can also apply policies to individual resources.

Folders



Say: The third level of the Google Cloud resource hierarchy is folders. Folders let you assign policies to resources at the level of granularity that you choose.

Folders let you assign policies to resources at the level of granularity that you choose. The resources in a folder inherit policies and permissions assigned to that folder. A folder can contain projects, other folders, or a combination of both.

Benefits of using the resource hierarchy to control access

01 Granular access control

02 Inheritance and propagation rules

03 Security and compliance

04 Strong visibility and auditing capabilities

Say: Now that you understand the structure of the Google Cloud resource hierarchy, let's explore some additional benefits of using it to control access to cloud resources.

1. First, the resource hierarchy provides **granular access control**, meaning you can assign roles and permissions at different levels of the hierarchy, such as at the folder, project, or individual resource level.

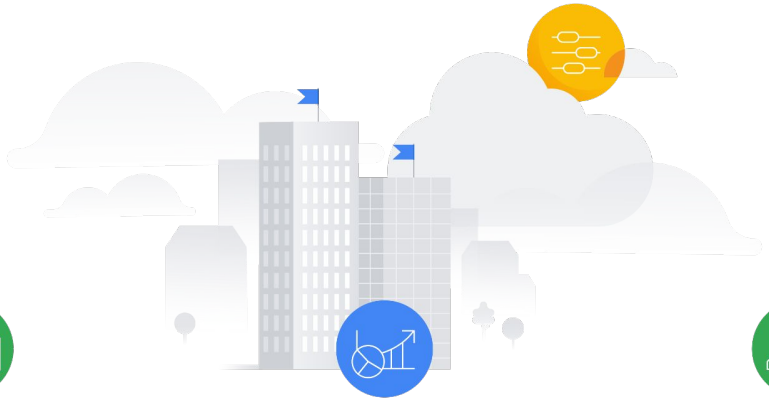
2. Second, because the resource hierarchy follows **inheritance and propagation rules**, permissions set at higher levels of the resource hierarchy are automatically inherited by lower-level resources. For example, if you grant a user access at the folder level, all projects and resources within that folder inherit those permissions by default. This inheritance simplifies access management and reduces the need for manual configuration at each individual resource level.

3. Third, the resource hierarchy enhances **security and compliance** through *least privilege* principles. By assigning access permissions at the appropriate level in the hierarchy, you can ensure that users *only* have the necessary privileges to perform their tasks. This reduces the risk of unauthorized access and helps maintain regulatory compliance.

4. Finally, the resource hierarchy provides **strong visibility and auditing capabilities**. You can track access permissions and changes across different levels of the hierarchy, which makes it easier to monitor and review access controls. This improves accountability and helps identify and address potential security issues.

Controlling cloud consumption

Organizations want to control cloud consumption for many reasons



Cost savings

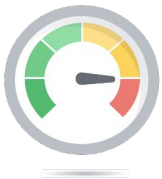
Increased visibility

Improved compliance

Say: Organizations want to control cloud consumption for many reasons. It could be about:

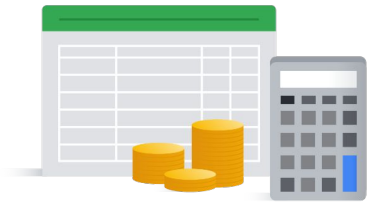
- **Cost savings** by ensuring they're not overspending on unnecessary resources.
- **Increased visibility** by providing a better understanding of how resources are being used and identifying areas to reduce costs.
- Or **improved compliance** by ensuring your cloud environment is compliant with industry regulations.

Google Cloud tools to help control cloud consumption



Resource quota policies

Set limits on the amount of resources that can be used by a project or user.



Budget threshold rules

Set alerts to be informed when your cloud costs exceed a certain threshold.

Cloud Billing reports

A reactive method to help track and understand what's already been spent.

Say: Google Cloud offers several tools to help control cloud consumption, including **resource quota policies**, **budget threshold rules,** and **Cloud Billing reports**. Let's define each of these terms.

- **Resource quota policies** let you set limits on the amount of resources that can be used by a project or user. They can help prevent overspending on cloud resources; therefore, they help you ensure that your cloud usage is within your budget.
- Then there are **budget threshold rules**, which let you set alerts to be informed when your cloud costs *exceed* a certain threshold. They can act as an early warning for potential cost overruns, and let you take corrective action before costs get out of control. Both resource quota policies and budget threshold rules are set in the Google Cloud console.
- And then there are **Cloud Billing reports.** Whereas resource quota policies and budget threshold rules provide proactive means to control cloud consumption, Cloud Billing reports offer a reactive method to help you track and understand what you've already spent on Google Cloud resources and provide ways to help optimize your costs. You can use Cloud Billing reports to monitor costs by exporting billing data to **BigQuery**. This means exporting usage and cost data to a BigQuery dataset, and then using the dataset for detailed analyses. You can also visualize data with tools like Looker Studio.

Optimizing costs through committed use discounts

Committed use discounts (CUDs)

Discounted prices in exchange for your commitment to use a minimum level of resources for a specific term.

Say: After analyzing how you're spending on cloud resources, you might realize that your organization can optimize costs through **committed use discounts (CUDs**).

If your workloads have predictable resource needs, you can purchase a Google Cloud commitment, which gives you discounted prices in exchange for your commitment to use a minimum level of resources for a specific term.

Quiz

Question

Which represents the lowest level in the Google Cloud resource hierarchy?

A. Folders
B. Projects
C. Organization node
D. Resources

Do: Read the question out loud. Ask the class to refrain from sharing their answers (either out loud or in the chat window) for about 10 seconds.

Say: Which represents the lowest level in the Google Cloud resource hierarchy?

A. Folders
B. Projects
C. Organization node
D. Resources

Quiz

Answer

Which represents the lowest level in the Google Cloud resource hierarchy?

A. Folders

B. Projects

C. Organization node

D. Resources 

Say: The correct answer is D.

A. Folders
   - Why this is the **incorrect** answer: Folders can contain other folders or projects, allowing further organization within your resource structure.

B. Projects
   - Why this is the **incorrect** answer: Projects are containers for resources, providing a logical grouping and a boundary for permissions and billing.

C. Organization node
   - Why this is the **incorrect** answer: This is the root of the hierarchy, representing your company or overarching entity.

D. Resources
   - Why this is the **correct** answer: Resources represent the lowest level in the Google Cloud resource hierarchy. Individual resources include things like virtual machine instances, Cloud Storage buckets, databases, and other cloud services.

Quiz

Question

Which feature lets you set limits on the amount of resources that can be used by a project or user?

A. Quota policies
B. Billing reports
C. Budget alerts
D. Committed use discounts

Do: Read the question out loud. Ask the class to refrain from sharing their answers (either out loud or in the chat window) for about 10 seconds.

Say: Which feature lets you set limits on the amount of resources that can be used by a project or user?

A. Quota policies
B. Billing reports
C. Budget alerts
D. Committed use discounts

Quiz

Answer

Which feature lets you set limits on the amount of resources that can be used by a project or user?

A. Quota policies 

B. Billing reports

C. Budget alerts

D. Committed use discounts

Say: The correct answer is A.

A. Quota policies
   - Why this is the **correct** answer: Quotas are the feature in Google Cloud that lets you enforce limits on the consumption of resources within a project or for a specific user.

B. Billing reports
   - Why this is the **incorrect** answer: Billing reports help you understand and track the costs associated with your resource usage but don't actively limit anything.

C. Budget alerts
   - Why this is the **incorrect** answer: Budget alerts can notify you when usage costs approach or exceed a specific threshold, providing helpful warnings but not actively restricting resource use.

D. Committed use discounts
   - Why this is the **incorrect** answer: This is a pricing model where you commit to using a certain amount of resources over a fixed period in exchange for discounted pricing; it doesn't provide resource usage limits.

Module 6

Scaling with Google Cloud Operations

Lessons

01 Financial governance and managing cloud costs

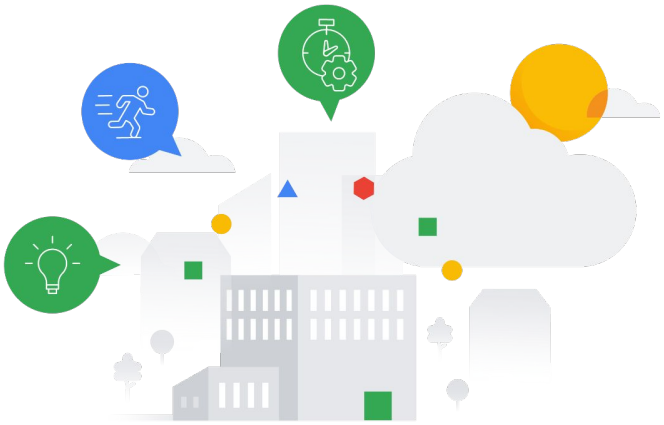
02 Operational excellence and reliability at scale

03 Sustainability with Google Cloud

Say: In today's rapidly evolving digital landscape, organizations use cloud technology increasingly to drive innovation, agility, and efficiency. However, harnessing the true power of the cloud requires a comprehensive understanding of operational excellence and reliability at scale.

Operational excellence and reliability

The ability to optimize operations and ensure uninterrupted service delivery, even when handling increasing workloads and complexities in the cloud.



Say: Operational excellence and reliability refers to the ability of organizations to optimize their operations and ensure uninterrupted service delivery, even as they handle increasing workloads and complexities in the cloud.

This includes designing robust infrastructure, establishing resilient processes, and employing proactive monitoring and response mechanisms.

Imagine a global ecommerce platform that experiences a sudden surge in traffic during a major sale event

Operational excellence

- Efficiently scaling the underlying infrastructure
- Automating resource provisioning
- Implementing load balancing mechanisms

Reliability at scale

- Minimizing downtime
- Employing fault-tolerant systems
- Employing disaster recovery strategies

Say: Imagine a global ecommerce platform that experiences a sudden surge in traffic during a major sale event. To meet the increased demand, the platform needs to scale its resources rapidly while ensuring uninterrupted service availability.

Operational excellence here involves efficiently scaling the underlying infrastructure, automating resource provisioning, and implementing load balancing mechanisms.

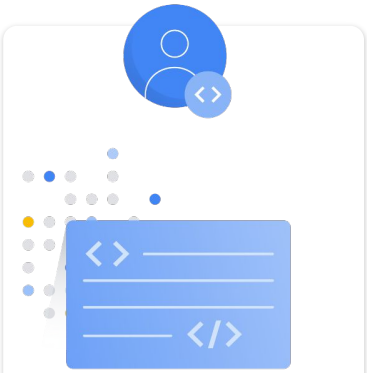
Reliability focuses on minimizing downtime, employing fault-tolerant systems, and employing disaster recovery strategies. By excelling in these areas, the ecommerce platform can handle the increased load seamlessly, deliver a consistently positive user experience, and avoid revenue loss or reputational damage.

In this section of the course, you explore:

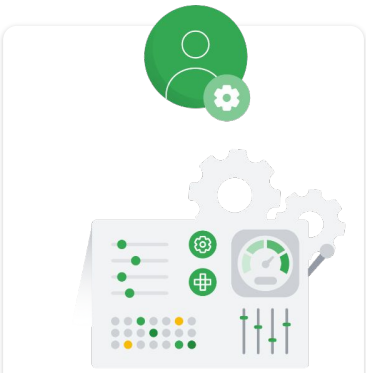
- Modernizing operations by using Google Cloud.
- Designing resilient infrastructure and processes.
- The fundamentals of cloud reliability.
- Google Cloud Customer Care.
- And the life of a support case.

The fundamentals of cloud reliability

Developers vs. operators



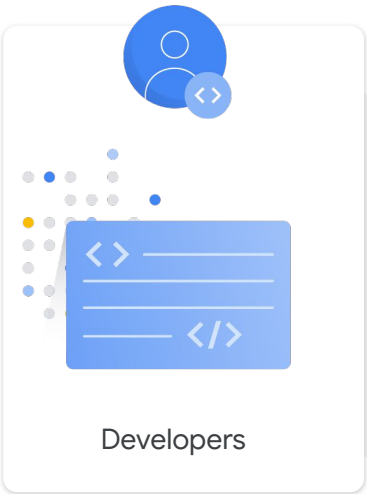
Developers write code for systems and applications.



Operators ensure those systems and applications operate reliably.

Say: Within any IT team, **developers** are responsible for writing code for systems and applications, and **operators** are responsible for ensuring that those systems and applications operate reliably.

Expectations of developers



- Are expected to be agile.
- Are often pushed to write and deploy code quickly.
- Aim to release new functions frequently.
- Increase core business value with new features.
- Release fixes fast.

Say: Developers are expected to be agile and are often pushed to write and deploy code quickly. Their aim is to release new functions frequently, increase core business value with new features, and release fixes fast for an overall better user experience.

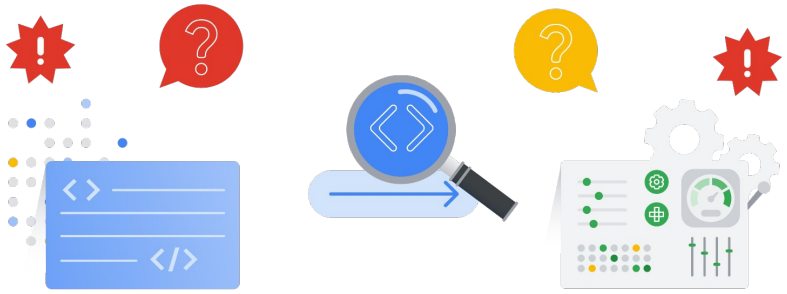
Expectations of operators

- Are expected to keep the system stable.
- Often prefer to work more slowly to ensure reliability and consistency.

Operators

Say: In contrast, operators are expected to keep the system stable, and so they often prefer to work more slowly to ensure reliability and consistency.

It can be difficult for either group to identify the source of problems and resolve them quickly

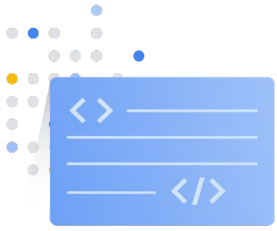


Traditionally, developers pushed their code to operators who often had little understanding of how the code runs in a live environment.

Say: Traditionally, developers pushed their code to operators who often had little understanding of how the code would run in a production or live environment.

When problems arise, it can be very difficult for either group to identify the source of the problem and resolve it quickly. Worse, accountability between the teams isn't always clear.

DevOps emphasizes collaboration and communication between development and operations teams



Developers

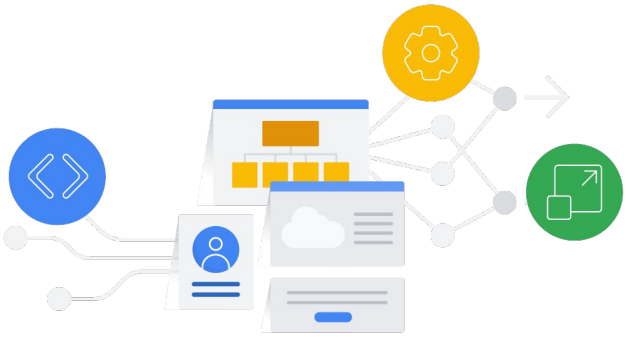
Operators

- Shared responsibility
- Automation
- Continuous improvement

Say: **DevOps** is a software development approach that emphasizes collaboration and communication between development and operations teams to enhance the efficiency, speed, and reliability of software delivery.

It aims to break down silos between these teams and foster a culture of shared responsibility, automation, and continuous improvement.

Site Reliability Engineering (SRE)



Ensures the reliability, availability, and efficiency of software systems and services deployed in the cloud.

Say: One particular concept within the DevOps framework is **Site Reliability Engineering**, or **SRE**, which ensures the reliability, availability, and efficiency of software systems and services deployed in the cloud.

SRE combines aspects of software engineering and operations to design, build, and maintain scalable and reliable infrastructure.

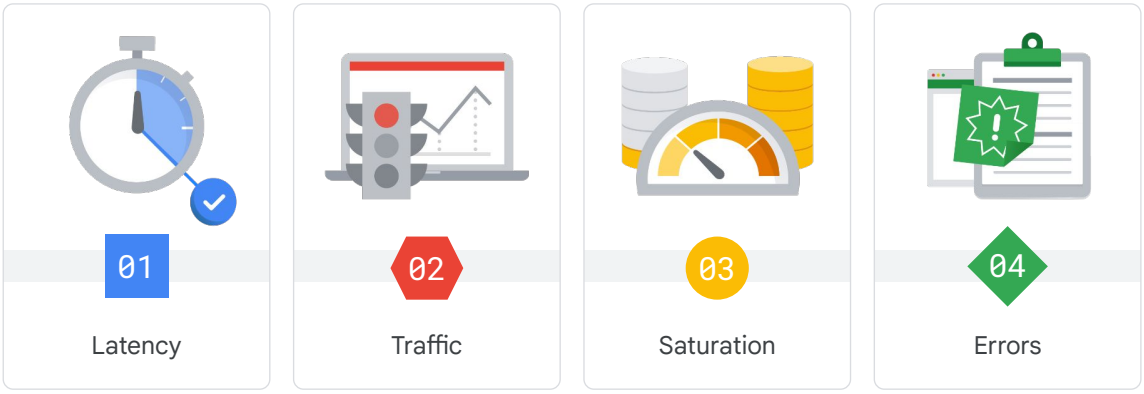
Monitoring is the foundation of product reliability

Monitoring

- Reveals what needs urgent attention.
- Shows trends in application usage patterns.
- Can yield better capacity planning.
- Help improve an application client's experience.

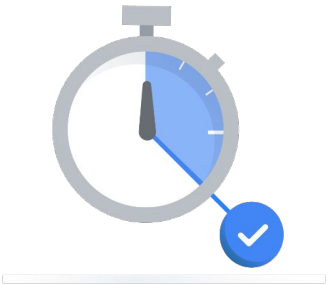
Say: Monitoring is the foundation of product reliability. It reveals what needs urgent attention and shows trends in application usage patterns, which can yield better capacity planning and generally help improve an application client's experience and lessen their pain.

Four golden signals measure a system's performance and reliability



Say: There are "four golden signals" that measure a system's performance and reliability. They are **latency**, **traffic**, **saturation**, and **errors**.

Latency



Latency measures how long it takes for a particular part of a system to return a result.

01 It directly affects the user experience.

02 Changes in latency could indicate emerging issues.

03 Its values might be tied to capacity demands.

04 It can be used to measure system improvements.

Say: **Latency** measures how long it takes for a particular part of a system to return a result.

Latency is important because:

- It directly affects the user experience.
- Changes could indicate emerging issues.
- Its values might be tied to capacity demands.
- And it can be used to measure system improvements.

Traffic

Traffic measures how many requests reach your system.

01 It's an indicator of current system demand.

02 Its historical trends are used for capacity planning.

03 It's a core measure when calculating infrastructure spend.

Say: **Traffic** measures how many requests reach your system.

Traffic is important because:

- It's an indicator of current system demand.
- Its historical trends are used for capacity planning.
- And it's a core measure when calculating infrastructure spend.

Saturation

Saturation measures how close to capacity a system is.

01 It's an indicator of how full the service is.

02 It focuses on the most constrained resources.

03 It’s frequently tied to degrading performance as capacity is reached.

Say: **Saturation** measures how close to capacity a system is. It’s important to note, though, that capacity is often a subjective measure that depends on the underlying service or application.

Saturation is important because:

- It's an indicator of how full the service is.
- It focuses on the most constrained resources.
- And it’s frequently tied to degrading performance as capacity is reached.

Errors

Errors are events that measure system failures or other issues.

01 They may indicate that something is failing.

02 They may indicate configuration or capacity issues.

03 They can indicate service level objective violations.

04 An error might mean it's time to send out an alert.

Say: And **errors** are events that measure system failures or other issues. Errors are often raised when a flaw, failure, or fault in a computer program or system causes it to produce incorrect or unexpected results, or behave in unintended ways.

Errors are important because they can indicate:

- Something is failing.
- Configuration or capacity issues.
- Service level objective violations.
- Or that it's time to send an alert.

Three main concepts in site reliability engineering

Service level indicators (SLIs)

Service level objectives (SLOs)

Service level agreements (SLAs)

Say: Three main concepts in site reliability engineering are **service-level indicators** (SLIs), **service-level objectives** (SLOs), and **service-level agreements** (SLAs). They are all types of targets set for a system's Four Golden Signal metrics.

SLIs

Service-level indicators

Measurements that show how well a system or service is performing.

- They're specific metrics like:
  - Response time
  - Error rate
  - Percentage uptime

Say: **Service level indicators** are measurements that show how well a system or service is performing. They're specific metrics like response time, error rate, or percentage uptime–which is the amount of time a system is available for use–that help us understand the system's behavior and performance.

SLOs

Service-level objectives

Goals that we set for a system's performance based on SLIs.

- They define what level of reliability or performance that we want to achieve.
  - → For example, an SLO might state that the system should be available for 99.9% of the time in a month.

Say: **Service level objectives** are the goals that we set for a system's performance based on SLIs. They define what level of reliability or performance that we want to achieve. For example, an SLO might state that the system should be available for 99.9% of the time in a month.

SLAs

Service-level agreements

Agreements between a cloud service provider and its customers.

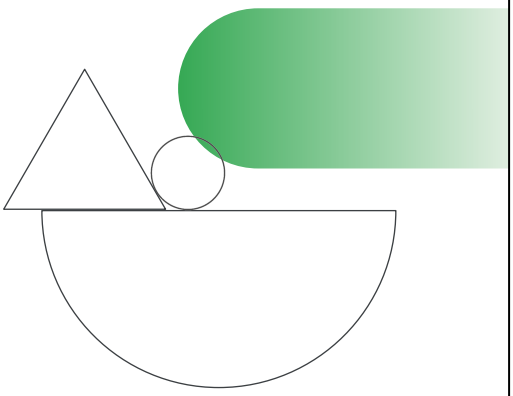
- They outline the promises and guarantees regarding the quality of service.
- They include the agreed-upon SLOs, performance metrics, uptime guarantees, and any penalties or remedies if the provider fails to meet those commitments.

Say: **Service level agreements** are agreements between a cloud service provider and its customers. They outline the promises and guarantees regarding the quality of service. SLAs include the agreed-upon SLOs, performance metrics, uptime guarantees, and any penalties or remedies if the provider fails to meet those commitments. This might include refunds or credits when the service has an outage that's longer than this agreement allows.

Activity

5 min | Class

For each of the scenarios that follow, identify whether they relate to Service Level Indicators (SLIs), Service Level Objectives (SLOs), or Service Level Agreements (SLAs).



Say: For each of the scenarios that follow, identify whether they relate to Service Level Indicators (SLIs), Service Level Objectives (SLOs), or Service Level Agreements (SLAs).

Is it an SLA, SLI, or SLO?

The website must be available for customers 99.95% of the time.

1

We had 2% downtime for our service last month.

2

The database query response time must be under 500 milliseconds.

3

If the customer support response time exceeds 24 hours, a discount will be applied.

4

The company will provide a partial refund to customers for each hour of service downtime.

5

If system errors exceed 5 per hour, an alert will be sent to the engineering team.

6

Say: Take a moment to read the six options on the screen.

Identify the two service-level objectives

The website must be available for customers 99.95% of the time. 1	We had 2% downtime for our service last month. 2	The database query response time must be under 500 milliseconds. 3
If the customer support response time exceeds 24 hours, a discount will be applied. 4	The company will provide a partial refund to customers for each hour of service downtime. 5	If system errors exceed 5 per hour, an alert will be sent to the engineering team. 6

Say: Which two represent **service-level objectives**. You'll recall that **SLOs** are the goals that we set for a system's performance, based on service level indicators.

Identify the two service-level objectives

The website must be available for customers 99.95% of the time. Service-level objective	We had 2% downtime for our service last month. 2	The database query response time must be under 500 milliseconds. Service-level objective
If the customer support response time exceeds 24 hours, a discount will be applied. 4	The company will provide a partial refund to customers for each hour of service downtime. 5	If system errors exceed 5 per hour, an alert will be sent to the engineering team. 6

Say: The correct answer is **options #1** and **#3.**

Identify the two service-level agreements

The website must be available for customers 99.95% of the time.

Service-level objective

We had 2% downtime for our service last month.

2

The database query response time must be under 500 milliseconds.

Service-level objective

If the customer support response time exceeds 24 hours, a discount will be applied.

4

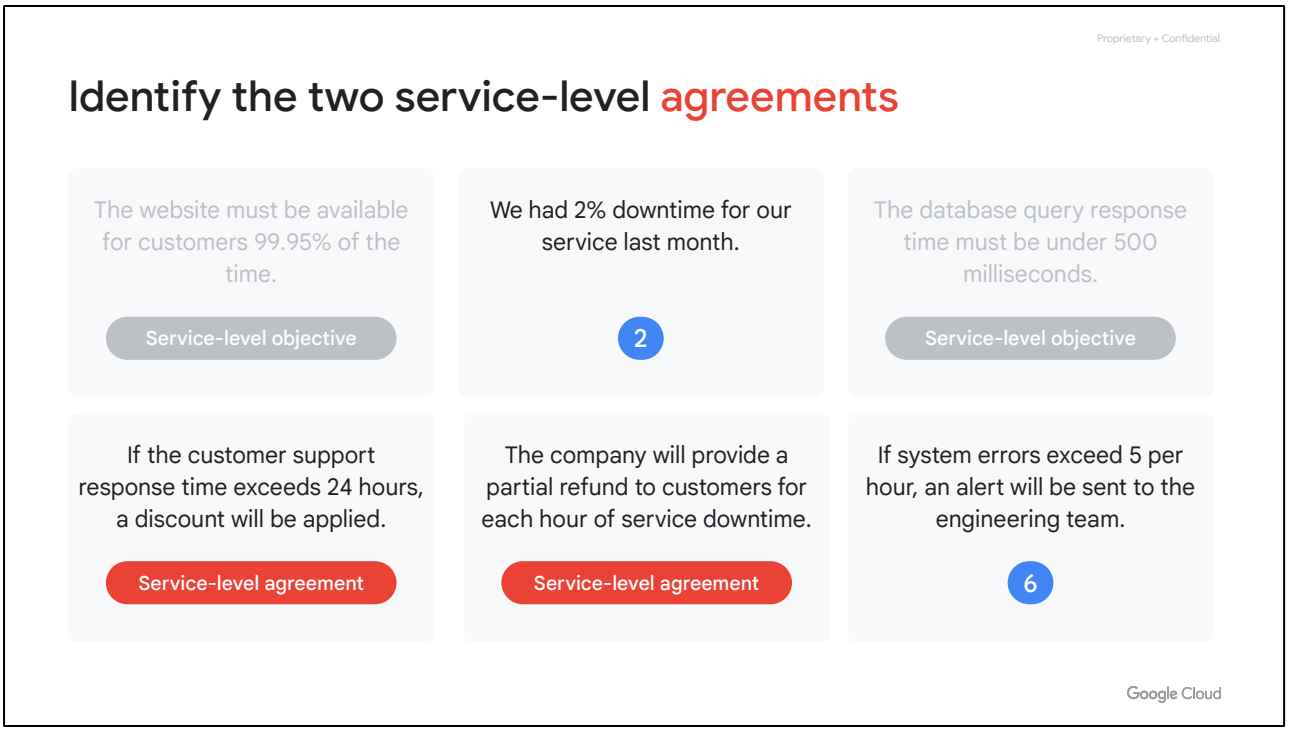
The company will provide a partial refund to customers for each hour of service downtime.

5

If system errors exceed 5 per hour, an alert will be sent to the engineering team.

6

Say: Next, which two options represent **service-level agreements**? SLAs are agreements between a cloud service provider and its customers.



Say: The correct answer is **options #4** and **#5.**

So where does that leave us? What's being described in **#2** and **#6**?

What are the remaining two?

The website must be available for customers 99.95% of the time.

Service-level objective

We had 2% downtime for our service last month.

2

The database query response time must be under 500 milliseconds.

Service-level objective

If the customer support response time exceeds 24 hours, a discount will be applied.

Service-level agreement

The company will provide a partial refund to customers for each hour of service downtime.

Service-level agreement

If system errors exceed 5 per hour, an alert will be sent to the engineering team.

6

Say: So where does that leave us? What's being described in **#2** and **#6**?

The service-level indicators

The website must be available for customers 99.95% of the time.	We had 2% downtime for our service last month.	The database query response time must be under 500 milliseconds.
Service-level objective	Service-level indicator	Service-level objective
If the customer support response time exceeds 24 hours, a discount will be applied.	The company will provide a partial refund to customers for each hour of service downtime.	If system errors exceed 5 per hour, an alert will be sent to the engineering team.
Service-level agreement	Service-level agreement	Service-level indicator

Say: The correct answer is **service-level indicators.** SLIs are measurements that show how well a system or service is performing.

Designing resilient infrastructure and processes

Cloud infrastructure and processes must be resilient, fault-tolerant, and scalable

High availability

The ability of a system to remain operational and accessible for users even if hardware or software failures occur

Disaster recovery

The process of restoring a system to a functional state after a major disruption or disaster

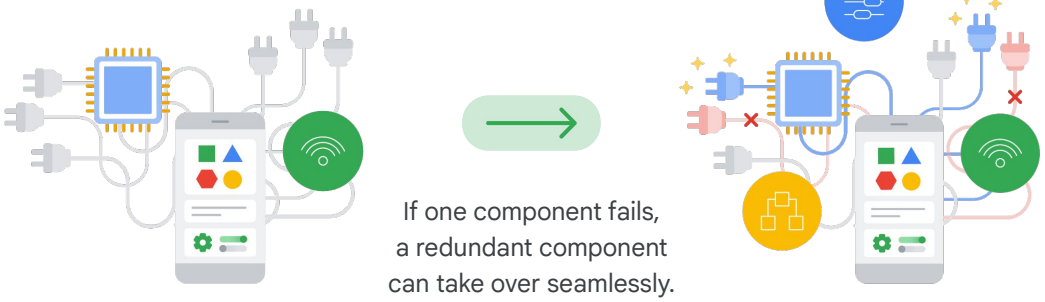
Say: When infrastructure and processes in a cloud environment are designed, they need to be resilient, fault-tolerant, and scalable, for high availability and disaster recovery.

- **High availability** refers to the ability of a system to remain operational and accessible for users even if hardware or software failures occur.
- **Disaster recovery** refers to the process of restoring a system to a functional state after a major disruption or disaster.

Let's explore some of the key design considerations and their significance in more detail.

Redundancy

Duplicating critical components or resources to provide backup alternatives.



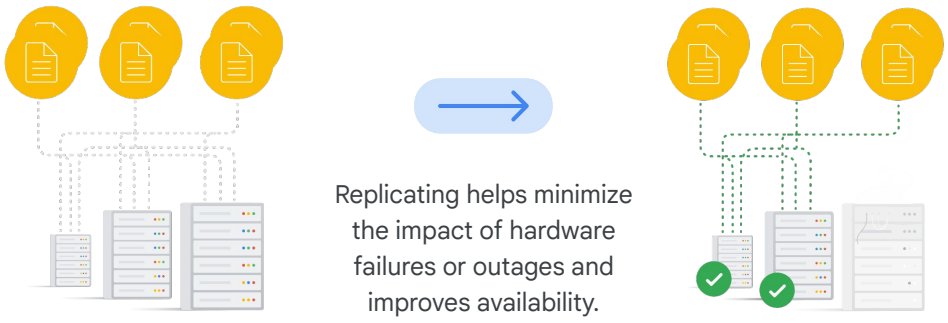
Say: **Redundancy** refers to duplicating critical components or resources to provide backup alternatives. Redundancy can be implemented at various levels, such as hardware, network, or application layers.

For example, having redundant power supplies, network switches, or load balancers ensures that if one fails, the redundant component takes over seamlessly.

Redundancy enhances system reliability and mitigates the impact of single points of failure.

Replication

Creating multiple copies of data or services and distributing them across different servers or locations.



Replicating helps minimize the impact of hardware failures or outages and improves availability.

Say: **Replication** involves creating multiple copies of data or services and distributing them across different servers or locations. It ensures redundancy and fault tolerance by allowing systems to continue functioning even if certain components or servers fail.

By replicating data across multiple servers, the impact of hardware failures or outages is minimized, and the availability of services is improved.

Regions

Cloud service providers offer multiple regions or data center locations spread across different geographic areas.

Ensures that if an entire region becomes unavailable, services can continue running from another region.

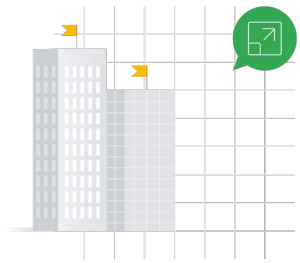
Say: Cloud service providers offer multiple **regions** or data center locations spread across different geographic areas.

By distributing resources across regions, businesses can ensure that if an entire region becomes unavailable due to natural disasters, network issues, or other incidents, their services can continue running from another region.

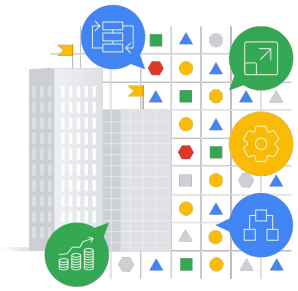
This approach improves resilience and reduces the risk of prolonged service interruptions.

Scalable infrastructure

Allows organizations to handle varying workloads and accommodate increased demand without compromising performance or availability.



Cloud technologies enable the dynamic allocation and deallocation of resources.

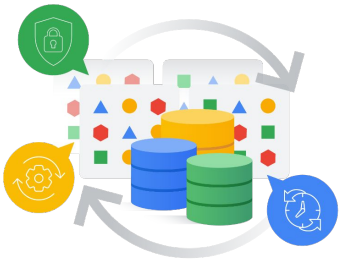


Say: Building a **scalable infrastructure** allows organizations to handle varying workloads and accommodate increased demand without compromising performance or availability.

Cloud technologies enable the dynamic allocation and deallocation of resources based on workload fluctuations. Autoscaling mechanisms can automatically adjust resource capacity to match demand, ensuring that services remain available and responsive during peak periods or sudden spikes in traffic.

Backups

Regular backups ensure that if data loss, hardware failures, or cyber-attacks occur, organizations can restore their systems to a previous state.



Cloud providers often offer backup services.

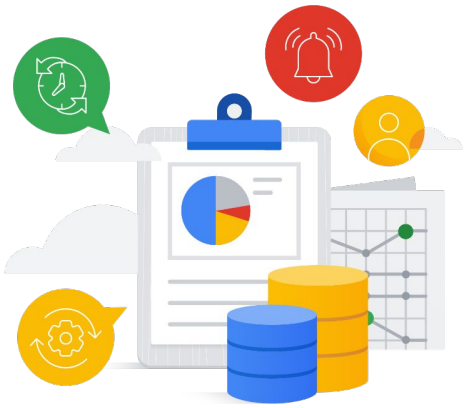
Backups should be stored in geographically separate locations.

Say: Regular **backups** of critical data and configurations are crucial to ensure that if data loss, hardware failures, or cyber-attacks occur, organizations can restore their systems to a previous state.

Cloud providers often offer backup services, and they let organizations automate backups, store them securely, and easily restore data when needed.

Backups should be stored in geographically separate locations to protect against regional outages or disasters.

Key design principles



- Improve high availability.
- Allow for rapid recovery from disasters or failures.
- Minimize downtime and data loss.
- It's important to test and validate these processes.
- Monitoring, alerting, and incident response mechanisms should be implemented.

Redundancy | Replication | Regions | Scalable infrastructure | Backups

Say: These measures improve high availability, allow for rapid recovery from disasters or failures, and minimize downtime and data loss.

It's important to regularly test and validate these processes to ensure that they function as expected during real-world incidents.

Also, monitoring, alerting, and incident response mechanisms should be implemented to identify and address issues promptly, further enhancing the overall resilience and availability of the cloud infrastructure.

Discussion

Designing resilient and scalable infrastructure for high availability and disaster recovery

- Think about the trade-offs involved in implementing redundancy, replication, and autoscaling in cloud environments.
- What factors should be considered when deciding on the appropriate level of redundancy and scalability for a specific application or service?

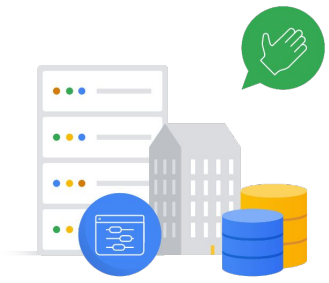
Say: Let's pause for a quick discussion around the importance of designing resilient and scalable infrastructure for high availability and disaster recovery.

Ask:

- Think about the trade-offs involved in implementing redundancy, replication, and autoscaling in cloud environments.
- What factors should be considered when deciding on the appropriate level of redundancy and scalability for a specific application or service?

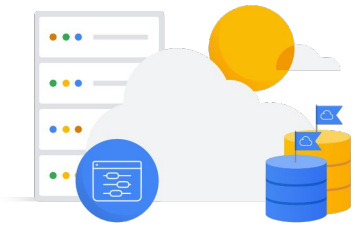
Modernizing operations using Google Cloud

Observability tools help gain insights into a system's performance, health, and behavior



On-premises

Physical check



Cloud

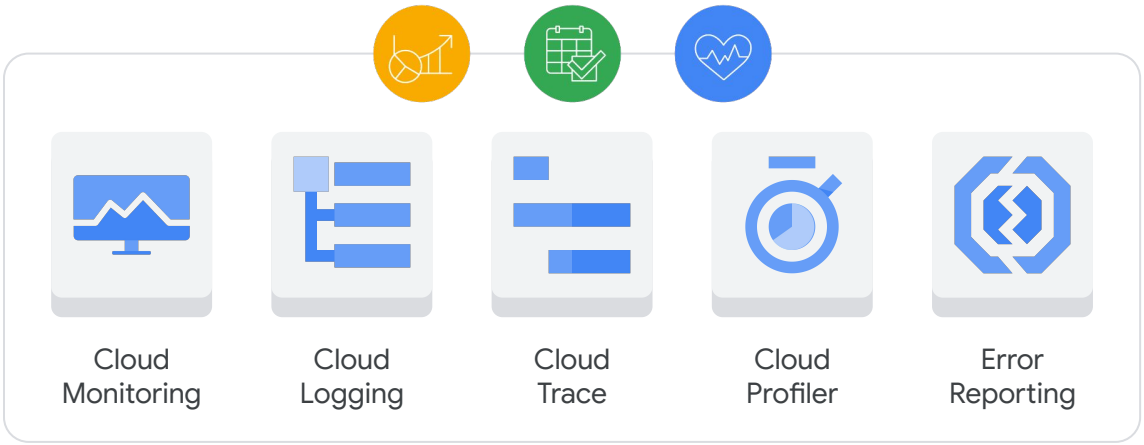
Observability tools

Say: If you've ever worked with on-premises environments, you know that you can physically touch the servers. If an application becomes unresponsive, someone can physically determine why that happened.

In the cloud though, the servers aren't yours—they belong to the cloud provider—and you can't physically inspect them. So the question becomes: how do you know what's happening with your server, database, or application?

The answer is: by using Google's integrated observability tools. Observability involves collecting, analyzing, and visualizing data from various sources within a system to gain insights into its performance, health, and behavior.

Google Cloud's operations suite



Say: To achieve this, Google Cloud offers an **operations suite**, which is a comprehensive set of monitoring, logging, and diagnostics tools. It offers a unified platform for managing and gaining insights into the performance, availability, and health of applications and infrastructure deployed on Google Cloud.

Let's look at some of the managed services that constitute the operations suite.

Cloud Monitoring

Cloud Monitoring

01 Provides a comprehensive view of cloud infrastructure and applications.

02 Collects metrics, logs, and traces from applications and infrastructure, and provides insights into their performance, health, and availability.

03 Lets you create alerting policies to notify when metrics, health check results, and uptime check results meet specified criteria.

Say: **Cloud Monitoring** provides a comprehensive view of your cloud infrastructure and applications. It collects metrics, logs, and traces from your applications and infrastructure, and provides you with insights into their performance, health, and availability.

It also lets you create alerting policies to notify you when metrics, health check results, and uptime check results meet specified criteria.

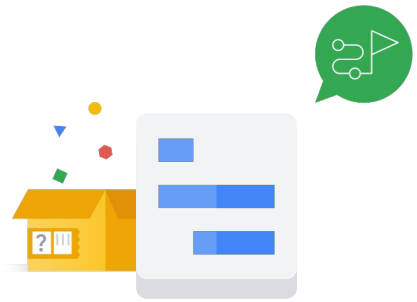
Cloud Logging

01 Collects and stores all application and infrastructure logs.

02 Real-time insights to help you troubleshoot issues, identify trends, and comply with regulations.

Say: **Cloud Logging** collects and stores all application and infrastructure logs. With real-time insights, you can use Cloud Logging to troubleshoot issues, identify trends, and comply with regulations.

Cloud Trace



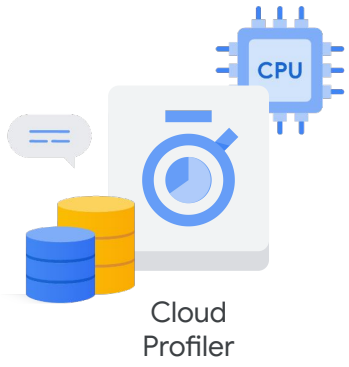
Cloud Trace

01 Helps identify performance bottlenecks in applications.

02 Collects latency data from applications, and provides insights into how they're performing.

Say: **Cloud Trace** helps identify performance bottlenecks in applications. It collects latency data from applications, and provides insights into how they're performing.

Cloud Profiler



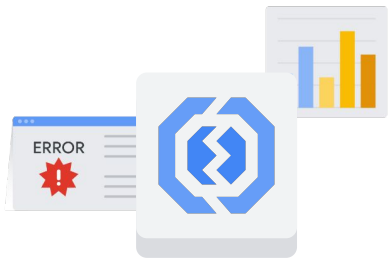
01 Identifies how much CPU power, memory, and other resources an application uses.

02 Continuously gathers CPU usage and memory-allocation information from production applications.

03 Provides insights into how applications are using resources.

Say: **Cloud Profiler** identifies how much CPU power, memory, and other resources an application uses. It continuously gathers CPU usage and memory-allocation information from production applications and provides insights into how applications are using resources.

Error Reporting



Error Reporting

01 Error Reporting counts, analyzes, and aggregates the crashes in running cloud services in real-time.

02 A centralized error management interface displays the results with sorting and filtering capabilities.

03 A dedicated view shows the error details: time chart, occurrences, affected user count, first- and last-seen dates.

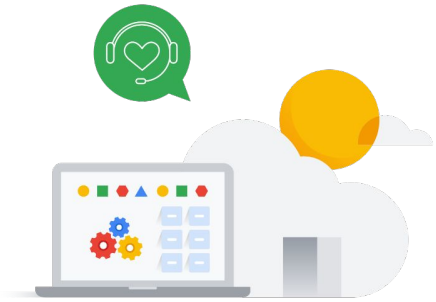
04 Error Reporting supports email and mobile notifications through its API.

Say: **Error Reporting** counts, analyzes, and aggregates the crashes in running cloud services in real-time. A centralized error management interface displays the results with sorting and filtering capabilities.

A dedicated view shows the error details: time chart, occurrences, affected user count, first- and last-seen dates, and a cleaned exception stack trace. Error Reporting supports email and mobile alerts notification through its API.

Google Cloud Customer Care

The four levels of Google Cloud Customer Care



- Basic
- Standard
- Enhanced
- Premium

Say: Any cloud adoption program can encounter challenges, so it's important to have an effective and efficient support plan from your cloud provider.

Google Cloud Customer Care can simplify and streamline your support experience with scalable and flexible services built with your business needs at the center.

There are four different service levels, which lets you choose the one that’s right for your organization.

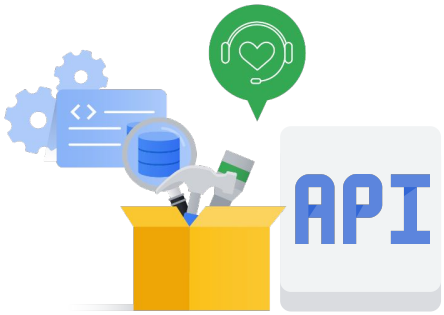
Basic Support

- Is free and included for all Google Cloud customers.
- Provides access to documentation, community support, Cloud Billing Support, and Active Assist recommendations.

Say: **Basic Support** is free and is included for all Google Cloud customers. It provides access to documentation, community support, Cloud Billing Support, and Active Assist recommendations.

Active Assist is the portfolio of tools used in Google Cloud to generate insights and recommendations to help you optimize your cloud projects.

Standard Support



Cloud
Support API

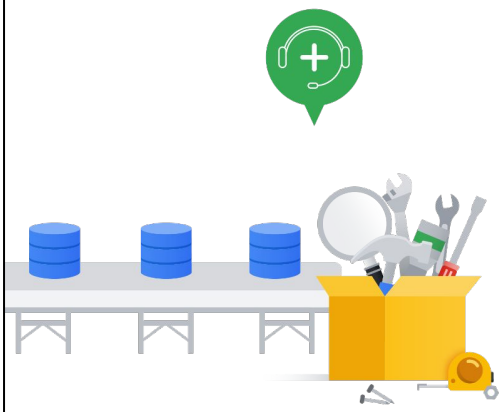
- Recommended for workloads under development.
- Provides unlimited access to tech support, which lets you troubleshoot, test, and explore.
- Offers unlimited individual access to English-speaking support representatives during working hours, five days a week.
- Provides access to the Cloud Support API.

Say: **Standard Support** is recommended for workloads under development. You can kickstart your cloud journey with unlimited access to tech support, which lets you troubleshoot, test, and explore.

It offers unlimited individual access to English-speaking support representatives during working hours, 5 days a week.

Standard support also provides access to the Cloud Support API, which lets you integrate Cloud Customer Care with your organization's customer relationship management (CRM) system.

Enhanced Support



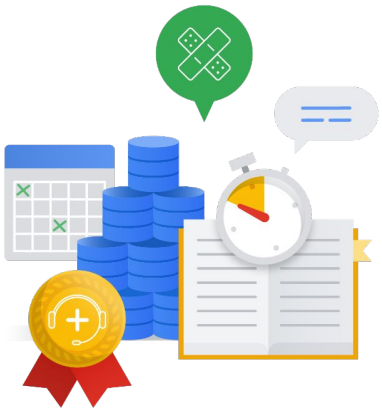
- Designed for workloads in production, with fast response times and additional services.
- Available 24/7 in a selection of languages, and initial response times are quicker than those provided by Standard Support.
- Offers technical support escalations and third-party technology support to help resolve multi-vendor issues.

Say: **Enhanced Support** is designed for workloads in production, with fast response times and additional services to optimize your experience with high-quality, robust support.

Support is available 24/7 in a selection of languages, and initial response times are quicker than those provided by Standard Support.

Enhanced Support also offers technical support escalations and third-party technology support to help you resolve multi-vendor issues.

Premium Support



- Designed for enterprises with critical workloads.
- Features the fastest response time, Customer Aware Support, and a dedicated Technical Account Manager.
- Includes credit for the Google Cloud Skills Boost training platform.
- Includes Event Management Service for planned peak events.
- Includes Operational Health Reviews.
- Includes Customer Aware Support.

Say: **Premium Support** is designed for enterprises with critical workloads. It features the fastest response time, Customer Aware Support, and a dedicated Technical Account Manager.

Our Premium Support level also offers:

- Credit for the Google Cloud Skills Boost training platform.
- An event management service for planned peak events, such as a product launch or major sales events.
- Operational health reviews to help you measure your progress and proactively address blockers to your goals with Google Cloud.
- And customer aware support, where Customer Care learns and maintains information about your architecture, partners, and Google Cloud projects. This information ensures that our support experts can resolve your cases promptly and efficiently.

Both the Enhanced and Premium support plans offer Value-Add Services that are available for additional purchase.

Google Cloud Customer Care

cloud.google.com/support

Say: You can learn more about the value-add services and all Google Cloud Customer Care support offerings at **cloud.google.com/support**.

The life of a support case

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure

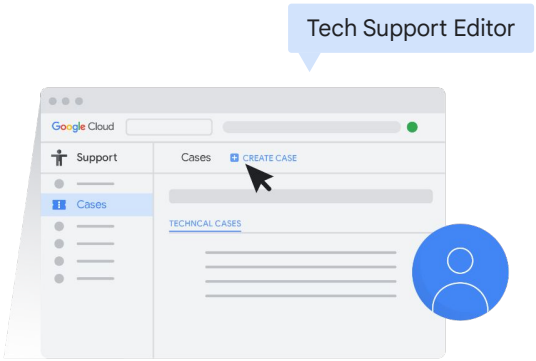
Say: Any Google Cloud customer on the Standard, Enhanced, or Premium Support plan can use the Google Cloud console to create and manage support cases.

Outside of filing a support case through the Google Cloud console, Customer Care Support also offers other contact options for live interactions with Support staff such as phone and video call support.

The life of a support case during the Google Cloud Customer Care process typically involves several stages and interactions between the customer and the support team. Here's an overview of the typical journey of a support case.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure



Google Cloud console

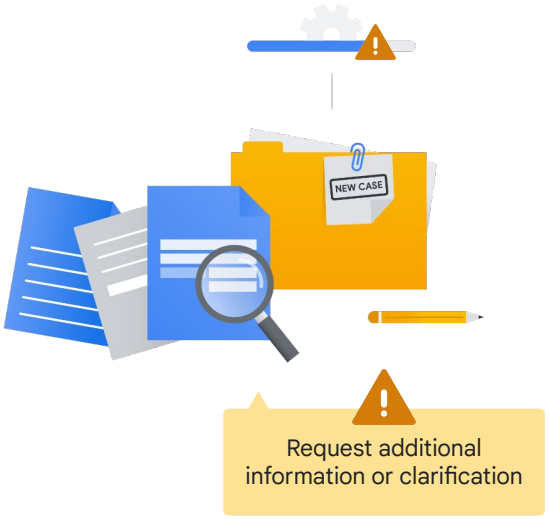
Say: First, the customer initiates the support request by **creating a case** in the Google Cloud Console. Only users who were assigned the Tech Support Editor role within an organization can do this.

The customer provides relevant details about the issue they are experiencing, including any error messages, logs, or steps to reproduce the problem.

It's important for the user to select a priority from P4, which means low impact, up to P1, which means critical impact, because this will influence response times from the Customer Care team.

The life of a support case

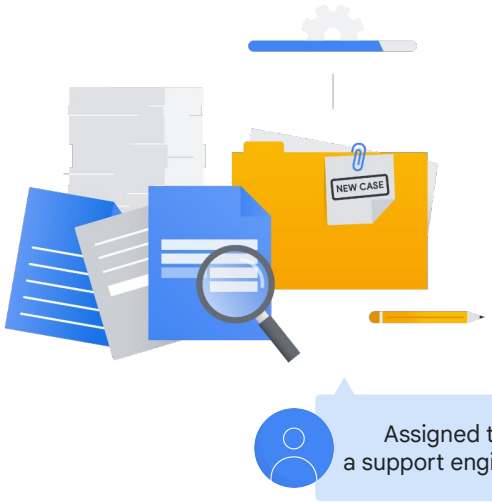
- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure



Say: After the case is created, it goes through a **triage** process. The team reviews the information provided by the customer to understand the problem and determine its severity and impact on the customer's business operations. The team might request additional information or clarification from the customer at this stage.

The life of a support case

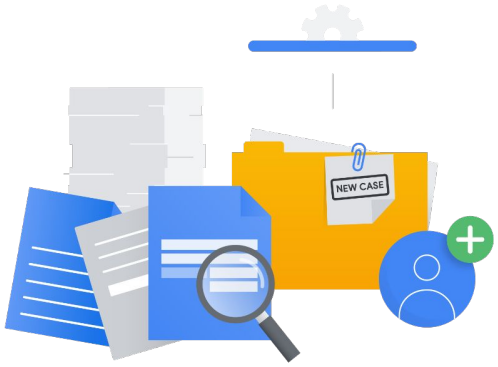
- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure



Say: In many cases, the Customer Care representative will resolve the case, but for more complex issues, the case is **assigned** to a support engineer with the appropriate level of expertise.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure

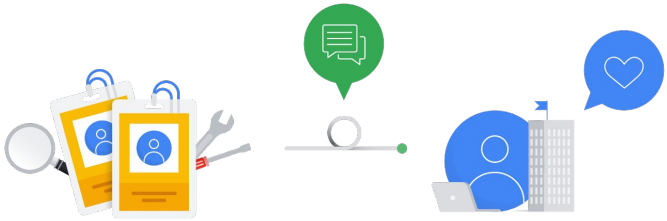


Analyze the provided information, review system logs, conduct diagnostic tests, and collaborate with other internal teams or experts.

Say: After the case is assigned, the team starts the **troubleshooting and investigation** process. They analyze the provided information, review system logs, and conduct various diagnostic tests to identify the root cause of the issue. Depending on the complexity of the problem, this stage might involve collaboration with other internal teams or experts.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- **Communication and updates**
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure



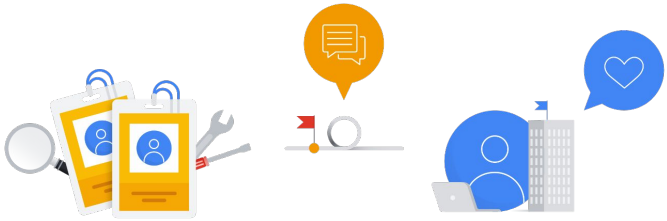
Customer Care team

Customer

Say: Throughout the investigation, the Customer Care team maintains regular **communication** with the customer. They provide **updates** on the progress, share findings, and request additional information or actions from the customer when needed.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- **Escalation**
- Resolution and mitigation
- Validation and testing
- Case closure



Customer Care team

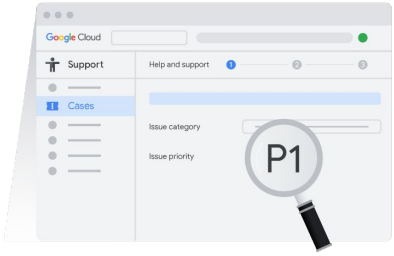
Customer

Say: **Escalation** is meant for flagging process breaks or for the rare occasion that a case is stuck because a customer and the Customer Care team aren't fully in sync, despite actively communicating the issue to determine the next steps.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure

Google Cloud console



High priority

Ensures that the case is assigned to the right resources as quickly as possible.

Say: However, it's important to note that escalation isn't always the best solution, and with high-impact issues, escalation might not make the case go faster. This is because escalation can disrupt the workflow of the Customer Care team and lead to delays in other cases.

The best solution for high-impact issues is to ensure that the case is set to the appropriate priority, ensuring that the case is assigned to the right resources as quickly as possible.

Escalation is a tool that can be used to regain traction on a stuck case. However, it's important to use escalation sparingly and only when it's absolutely necessary.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure

Customer Care team

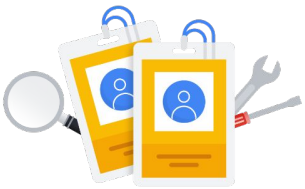
Say: When the root cause is identified, the team works on **resolving** the issue or providing a mitigation plan.

They might provide the customer with step-by-step instructions, configuration changes, or workaround suggestions to address the problem.

In some cases, they might consult the issue with higher-level support or engineering teams for further assistance. The Customer Care team might also need to submit a feature request to the Google Cloud engineering team.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- **Validation and testing**
- Case closure



Customer Care team

Perform specific tests

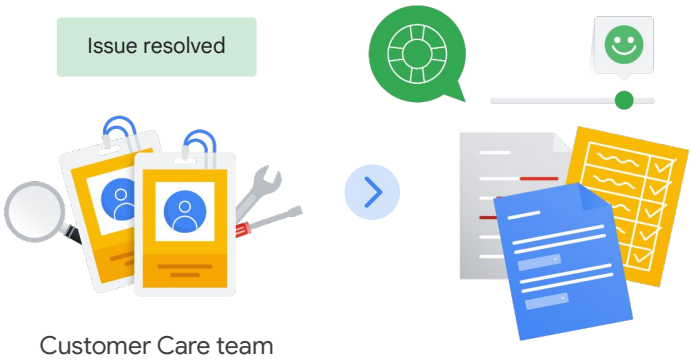
Provide feedback

Say: After implementing the resolution or mitigation plan, the Customer Care team collaborates with the customer to **validate** the effectiveness of the solution.

They might request the customer to perform specific tests or provide feedback on the outcome. This step ensures that the problem is fully resolved and meets the customer's expectations.

The life of a support case

- Case creation
- Case triage
- Case assignment
- Troubleshooting and investigation
- Communication and updates
- Escalation
- Resolution and mitigation
- Validation and testing
- Case closure



Say: When the customer confirms that the issue is resolved, the support case is **closed**.

The team provides a summary of the resolution, documents the steps taken, and ensures that the customer is satisfied with the outcome. If needed, they might also offer recommendations for preventive measures or future best practices to avoid similar issues.

The customer also receives a feedback survey, so the support team can learn what they did well and what needs improvement.

Throughout the entire lifecycle of the support case, Google Cloud's Customer Care team aims to provide timely and effective assistance to the customer. They prioritize customer satisfaction, responsiveness, and strive to address the possible technical challenges faced by customers when they use Google Cloud services.

Quiz

Question

Google Cloud's operations suite provides a comprehensive set of monitoring, logging, and diagnostics tools. Which tool collects latency data from applications and provides insights into how they're performing?

A. Cloud Profiler

B. Cloud Monitoring

C. Cloud Trace

D. Cloud Logging

Do: Read the question out loud. Ask the class to refrain from sharing their answers (either out loud or in the chat window) for about 10 seconds.

Say: Google Cloud's operations suite provides a comprehensive set of monitoring, logging, and diagnostics tools. Which tool collects latency data from applications and provides insights into how they're performing?

A. Cloud Profiler
B. Cloud Monitoring
C. Cloud Trace
D. Cloud Logging

Quiz

Answer

Google Cloud's operations suite provides a comprehensive set of monitoring, logging, and diagnostics tools. Which tool collects latency data from applications and provides insights into how they're performing?

A. Cloud Profiler

B. Cloud Monitoring

C. Cloud Trace 

D. Cloud Logging

Say: The correct answer is C.

A. Cloud Profiler
- Why this is the **incorrect** answer: This tool analyzes the resource usage (CPU, memory) of your applications to help identify performance issues related to code efficiency rather than network or system latency.

B. Cloud Monitoring
- Why this is the **incorrect** answer: Broader in scope, Cloud Monitoring collects various metrics, including some latency-related ones. However, Cloud Trace provides deeper analysis and visualization specifically geared towards understanding latency issues.

C. Cloud Trace
- Why this is the **correct** answer: Cloud Trace is the tool within Google Cloud's operations suite that specializes in collecting latency data and analyzing application performance. It shows exactly how long it takes for requests to move through various parts of your system, helping identify bottlenecks and areas for optimization.

D. Cloud Logging
- Why this is the **incorrect** answer: Cloud Logging gathers and stores logs from your applications and Google Cloud services. While analyzing these logs could yield latency insights, it's not the primary purpose of the tool.

Quiz

Question

Which Google Cloud Customer Care support level is designed for enterprises with critical workloads and features the fastest response time?

A. Standard Support

B. Premium Support

C. Basic Support

D. Enhanced Support

Do: Read the question out loud. Ask the class to refrain from sharing their answers (either out loud or in the chat window) for about 10 seconds.

Say: Which Google Cloud Customer Care support level is designed for enterprises with critical workloads and features the fastest response time?

A. Standard Support
B. Premium Support
C. Basic Support
D. Enhanced Support

Quiz

Answer

Which Google Cloud Customer Care support level is designed for enterprises with critical workloads and features the fastest response time?

A. Standard Support

B. Premium Support 

C. Basic Support

D. Enhanced Support

Say: The correct answer is B.

A. Standard Support
- Why this is the **incorrect** answer: This offers less urgent response times (typically within a business day) and is suitable for non-critical workloads.

B. Premium Support
- Why this is the **correct** answer: Premium Support is the Google Cloud Customer Care support level designed for large enterprises with mission-critical workloads. It offers the fastest response times (15-minute target for the most urgent issues), proactive guidance, and a Technical Account Manager for direct personalized support.

C. Basic Support
- Why this is the **incorrect** answer: This level includes access to documentation and community forums, but no direct support from Google engineers.

D. Enhanced Support
- Why this is the **incorrect** answer: Enhanced Support sits between Standard and Premium, offering shorter response times than Standard but less comprehensive support than Premium.

Module 6

Scaling with Google Cloud Operations

Lessons

01 Financial governance and managing cloud costs

02 Operational excellence and reliability at scale

03 Sustainability with Google Cloud

Say: As we get closer to the end of this Cloud Digital Leader training, where you've explored how cloud computing can help transform the way you do business, it's important that we underscore our technology efforts at Google with our commitment to the environment and sustainability.

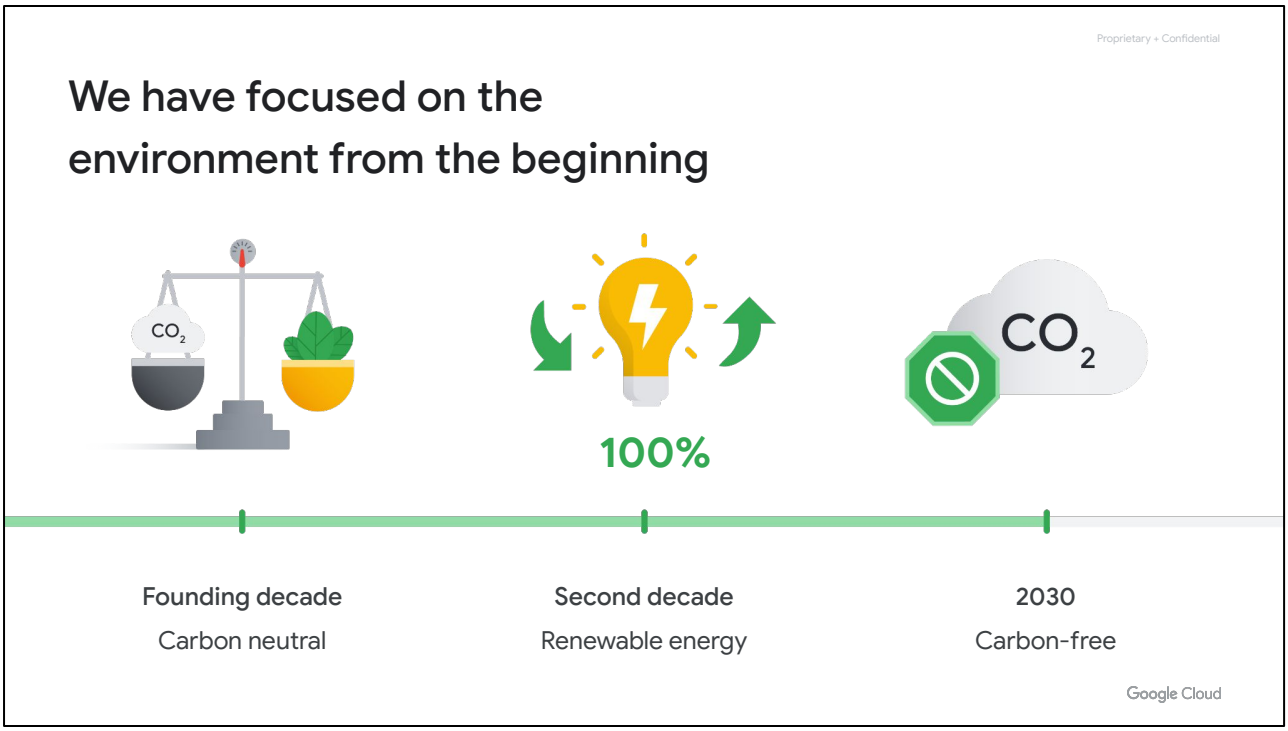
Just like our customers, Google tries to take care of the planet

Google's data centers were the first to achieve ISO 14001 certification

Our customers

Say: Just like our customers, Google is trying to take care of the planet. We understand that Google Cloud customers have environmental goals of their own, and running their workloads on Google Cloud can be a part of meeting those goals.

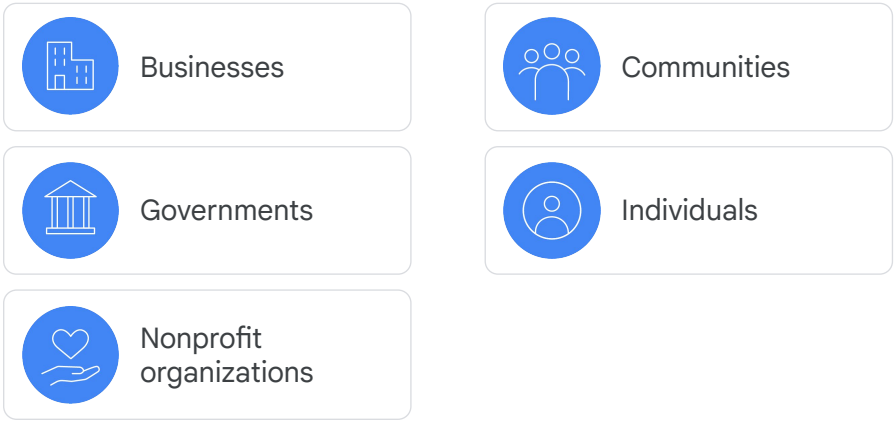
Therefore, it's useful to note that Google's data centers were the first to achieve **ISO 14001 certification**, which is a standard that outlines a framework for an organization to enhance its environmental performance through improving resource efficiency and reducing waste.



Say: As you heard from Sundar in the last video, Google has focused on the environment from the beginning.

In our founding decade, Google became the first major company to be **carbon neutral**. In our second decade, we were the first company to achieve **100% renewable energy**. And by 2030, we aim to be the first major company to operate completely **carbon free**.

We meet the challenges of climate change and the need for resource efficiency by working to empower everyone



Say: We meet the challenges posed by climate change and the need for resource efficiency by working to empower everyone—businesses, governments, nonprofit organizations, communities, and individuals—to use Google technology to create a more sustainable world.

We remain committed to sustainability and improving the health of our planet

Google Cloud

Say: At Google, we remain committed to sustainability and continue to lead and encourage others, like Kaluza, to join us in improving the health of our planet.

Quiz

Question

What sustainability goal does Google aim to achieve by the year 2030?

A. To be the first major company to achieve 100% renewable energy.

B. To be the first major company to be carbon neutral.

C. To be the first major company to run its own wind farm.

D. To be the first major company to operate completely carbon free.

Do: Read the question out loud. Ask the class to refrain from sharing their answers (either out loud or in the chat window) for about 10 seconds.

Say: What sustainability goal does Google aim to achieve by the year 2030?

A. To be the first major company to achieve 100% renewable energy.
B. To be the first major company to be carbon neutral.
C. To be the first major company to run its own wind farm.
D. To be the first major company to operate completely carbon free.

Quiz

Answer

What sustainability goal does Google aim to achieve by the year 2030?

A. To be the first major company to achieve 100% renewable energy.

B. To be the first major company to be carbon neutral.

C. To be the first major company to run its own wind farm.

D. To be the first major company to operate completely carbon-free.

Say: The correct answer is D.

A. To be the first major company to achieve 100% renewable energy.
- Why this is the **incorrect** answer: Google already achieved this in 2017, matching their annual global electricity consumption with 100% renewable energy purchases. Operating carbon-free 24/7 is an even more ambitious step.

B. To be the first major company to be carbon neutral.
- Why this is the **incorrect** answer: Google achieved carbon neutrality in 2007 and continues to offset its legacy carbon emissions. But their 2030 goal is about carbon-free operation, not neutrality through offsets.

C. To be the first major company to run its own wind farm.
- Why this is the **incorrect** answer: Google invests in renewable energy projects, including wind farms, but the broader goal isn't solely about ownership, but rather ensuring every hour of operation is matched with directly-sourced carbon-free energy.

D. To be the first major company to operate completely carbon-free.
- Why this is the **correct** answer: Google aims to eliminate the use of fossil fuels completely by 2030 across their data centers and office campuses, ensuring all operations are powered by renewable sources like wind and solar 24 hours a day, 7 days a week.

Quiz

Question

Google's data centers were the first to achieve ISO 14001 certification. What is this standard's purpose?

A. It's a framework for an organization to enhance its environmental performance through improving resource efficiency and reducing waste.

B. It's a framework for identifying, predicting, and evaluating the environmental impacts of a proposed project.

C. It's a framework for carbon footprinting that calculates the total amount of greenhouse gas emissions associated with a product, service, or organization.

D. It's a framework for sustainable procurement, which is the process of purchasing goods and services in a way that minimizes environmental and social impacts.

Do: Read the question out loud. Ask the class to refrain from sharing their answers (either out loud or in the chat window) for about 10 seconds.

Say: Google's data centers were the first to achieve ISO 14001 certification. What is this standard's purpose?

A. It's a framework for an organization to enhance its environmental performance through improving resource efficiency and reducing waste.
B. It's a framework for identifying, predicting, and evaluating the environmental impacts of a proposed project.
C. It's a framework for carbon footprinting that calculates the total amount of greenhouse gas emissions associated with a product, service, or organization.
D. It's a framework for sustainable procurement, which is the process of purchasing goods and services in a way that minimizes environmental and social impacts.

Quiz

Answer

Google's data centers were the first to achieve ISO 14001 certification. What is this standard's purpose?

A. It's a framework for an organization to enhance its environmental performance through improving resource efficiency and reducing waste.

B. It's a framework for identifying, predicting, and evaluating the environmental impacts of a proposed project.

C. It's a framework for carbon footprinting that calculates the total amount of greenhouse gas emissions associated with a product, service, or organization.

D. It's a framework for sustainable procurement, which is the process of purchasing goods and services in a way that minimizes environmental and social impacts.

Say: The correct answer is A.

A. It's a framework for an organization to enhance its environmental performance through improving resource efficiency and reducing waste.
   - Why this is the **correct** answer: This is the correct description of the ISO 14001 standard.

B. It's a framework for identifying, predicting, and evaluating the environmental impacts of a proposed project.
   - Why this is the **incorrect** answer: This description is closer to an Environmental Impact Assessment (EIA). ISO 14001 doesn't focus on individual projects but the overall environmental management system of an organization.

C. It's a framework for carbon footprinting that calculates the total amount of greenhouse gas emissions associated with a product, service, or organization.
   - Why this is the **incorrect** answer: While reducing carbon emissions is a likely positive outcome of an ISO 14001-driven system, the standard itself is broader than just carbon footprinting.

D. It's a framework for sustainable procurement, which is the process of purchasing goods and services in a way that minimizes environmental and social impacts.
   - Why this is the **incorrect** answer: Sustainable procurement could be one important area considered within an organization's ISO 14001 environmental management system, but the standard itself is broader.