Google Cloud

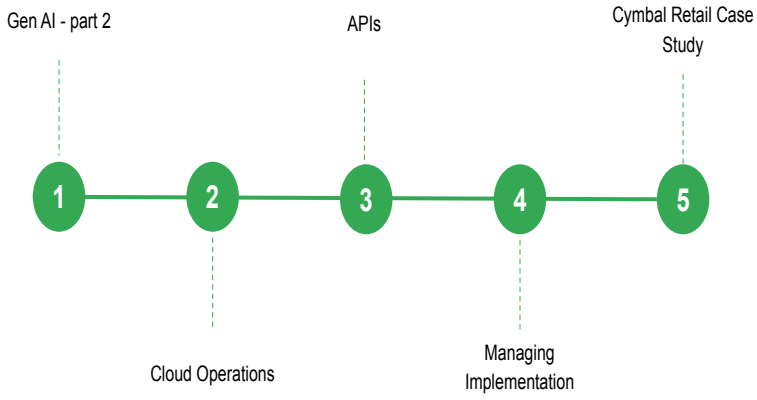
Professional Cloud Architect

Preparing for Professional Cloud Architect Journey for AWS Professionals

Plan:

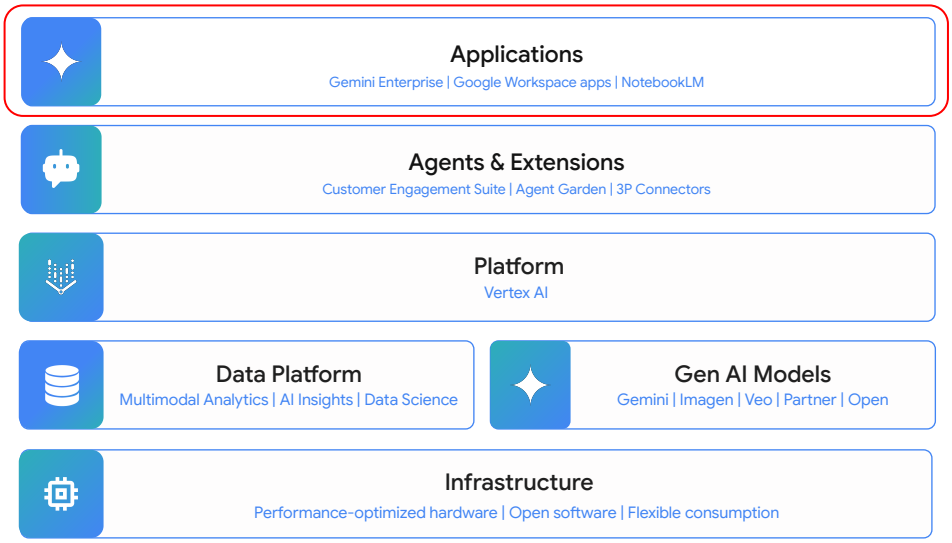
- 6 min explanation and [DEMO] VPC-S.C.
- 6 min explanation and [DEMO] Security Center
- 7 min explanation and [DEMO] IAP + Access Levels
- 6 min explanation and [DEMO] Log router
- 7 min: IaC:
  - Mostly mention about the preferred aproach
  - Terraform (Deployment Manager is not preferred anymore)
  - Show "Google Cloud Setup" in Cloud Console -> terraform code
  - Show github automation resources (from slides, mostly FAST)
- 6 mins HA "game"
- ~10 mins: [CASE STUDY] Mountkirk Games

Session 7 topics



Gen AI - Part 2

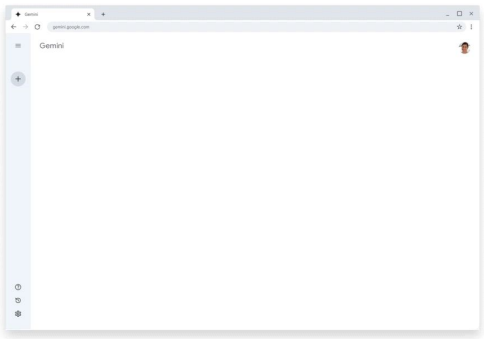
Google's Unified AI technology stack



Our final stop in the technology stack is the Applications layer, which includes Gemini, Google Workspace apps, and NotebookLM

Gemini app

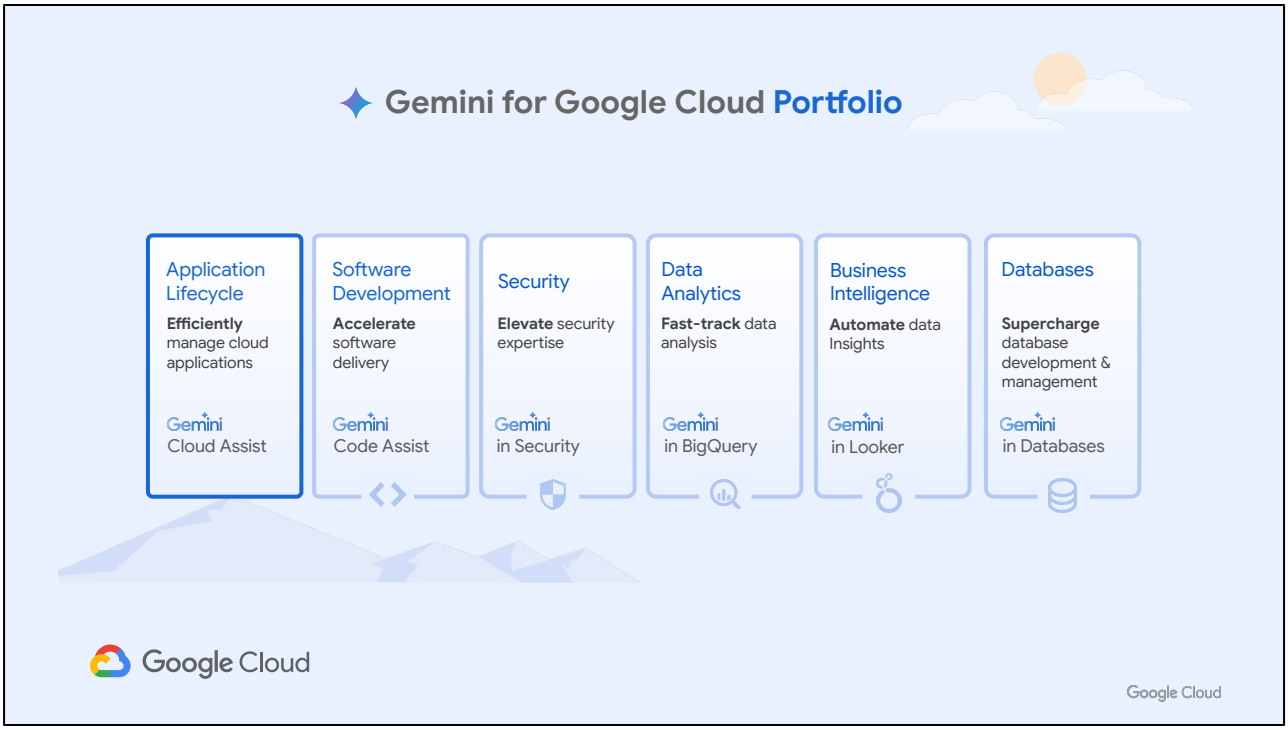
Chat with Gemini to kickstart brainstorming and planning, or tackle complex projects like research and coding.



Google Cloud

First, the Gemini app. You can use conversational chat with Gemini for brainstorming, planning, or tackle complex tasks like coding and research. You can use images, videos, text, and code in your input, and get all of those modes as outputs as well.

- Show https://gemini.google/release-notes/



1. Cloud Assist

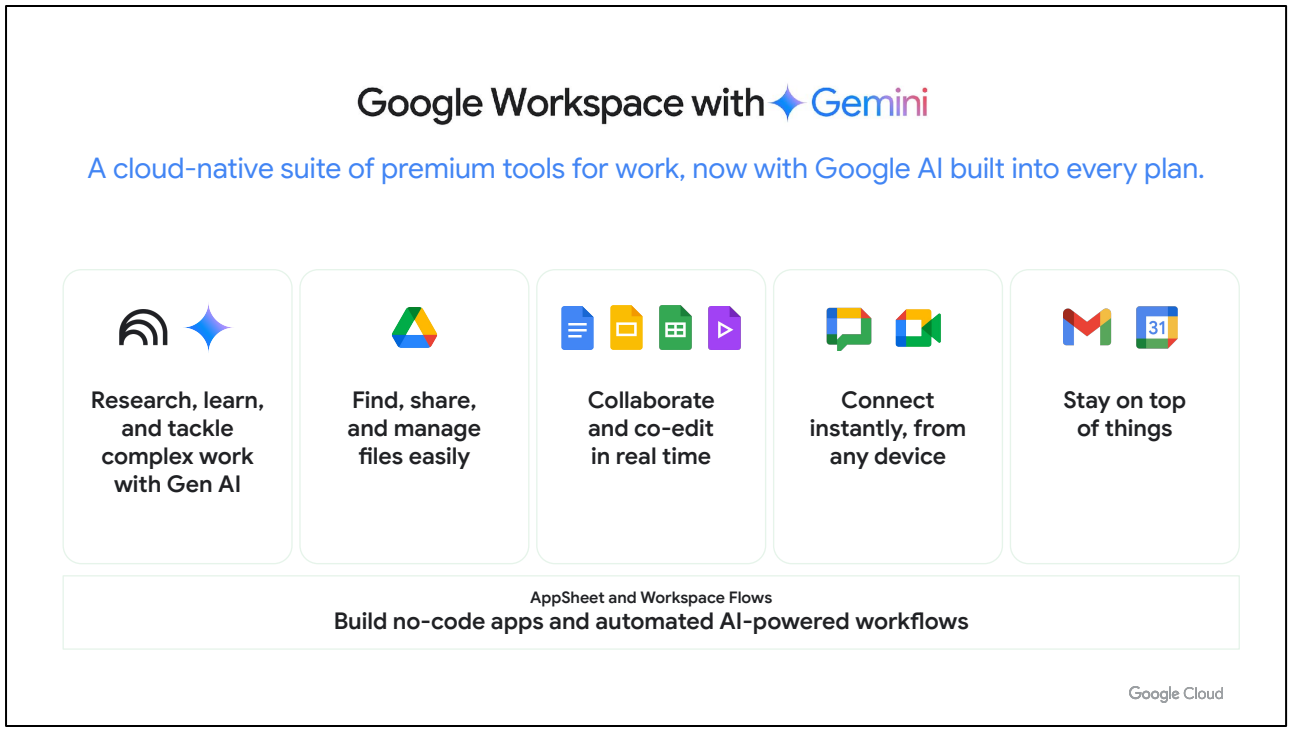
As we apply these GenAI capabilities across Google Cloud we raise the productivity bar for all GCP users. We create new experiences across our cloud offerings that helps accelerate the velocity of your IT. With that accelerated IT delivery, you can make your internal processes more efficient. You can also achieve outcomes that drive business growth, become more competitive, and create new experiences for your customers. These outcomes are possible because Google Cloud offers the first GenAI-powered cloud that has integrated assistive and collaborative capabilities across all of our cloud services.

By integrating AI-powered assistance with Google Developer Cloud we make your developers and operators more productive so that you can innovate with faster software delivery lifecycle and deliver new functional capabilities to your end customers.

By integrating AI-powered assistance across our Data and AI cloud,

we make your data users and data systems more productive and efficient. This assistance across data and resource management helps your data scientists aggregate data faster, reach decisions faster and deliver new personalized experiences to your users and customers.

Similarly, with Gemini capabilities infused across our Security Cloud, we help you reduce the risk of threat overload, grapple with toilsome tools that help with forensics and keep your IT environments secure.

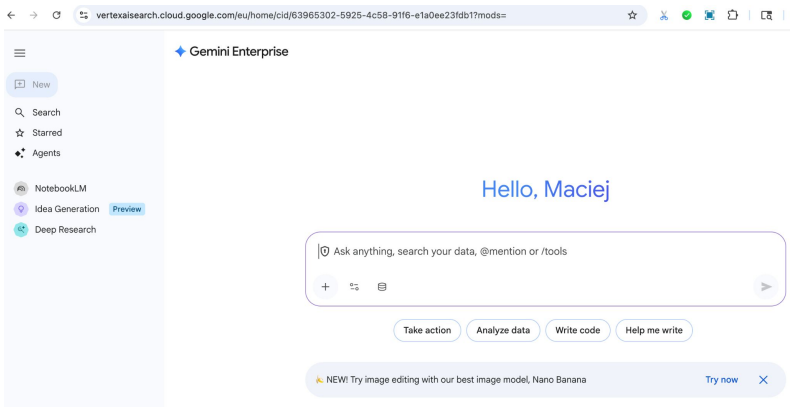


And finally, Gemini is there in all of Google Workspace applications. In gmail, it can help you write a formal letter, in Docs it will make grammar suggestions, in Slides it can generate a slide based on your prompt and more.
To sum up: Gemini is being injected everywhere nowadays, so we're kind of evolving from a central, chat-like application to having some helpers here and there to support us with common, daily tasks that we usually perform at work.

– DEMO sth here!!! Gmail / docs / slides etc. You might get asked about those productivity improvements on the exam!

Gemini Enterprise

Unlock enterprise expertise for employees with agents that bring together Gemini's advanced reasoning, Google-quality search, and enterprise data.

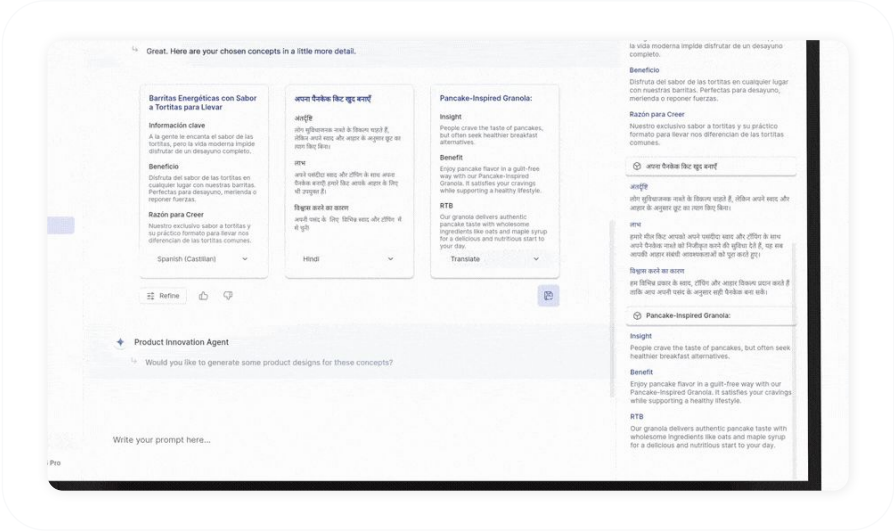


– Agentspace is being replaced by Gemini Enterprise plans as of late 2025
Google Agentspace is now part of Gemini Enterprise. The conversational AI and agent creation and orchestration technology behind Agentspace is now powering the core functionalities of the Gemini Enterprise platform.

****** SHORT DEMO to explain Agentspace is an application deployed and configured via GCP console first, and then exposed to end users *****

It helps unlock enterprise expertise for employees. It integrates Gemini's powerful reasoning and Google-quality search capabilities with your organization's data. It allows employees to get work done by searching across all your systems with the highest levels of compliance and data protection

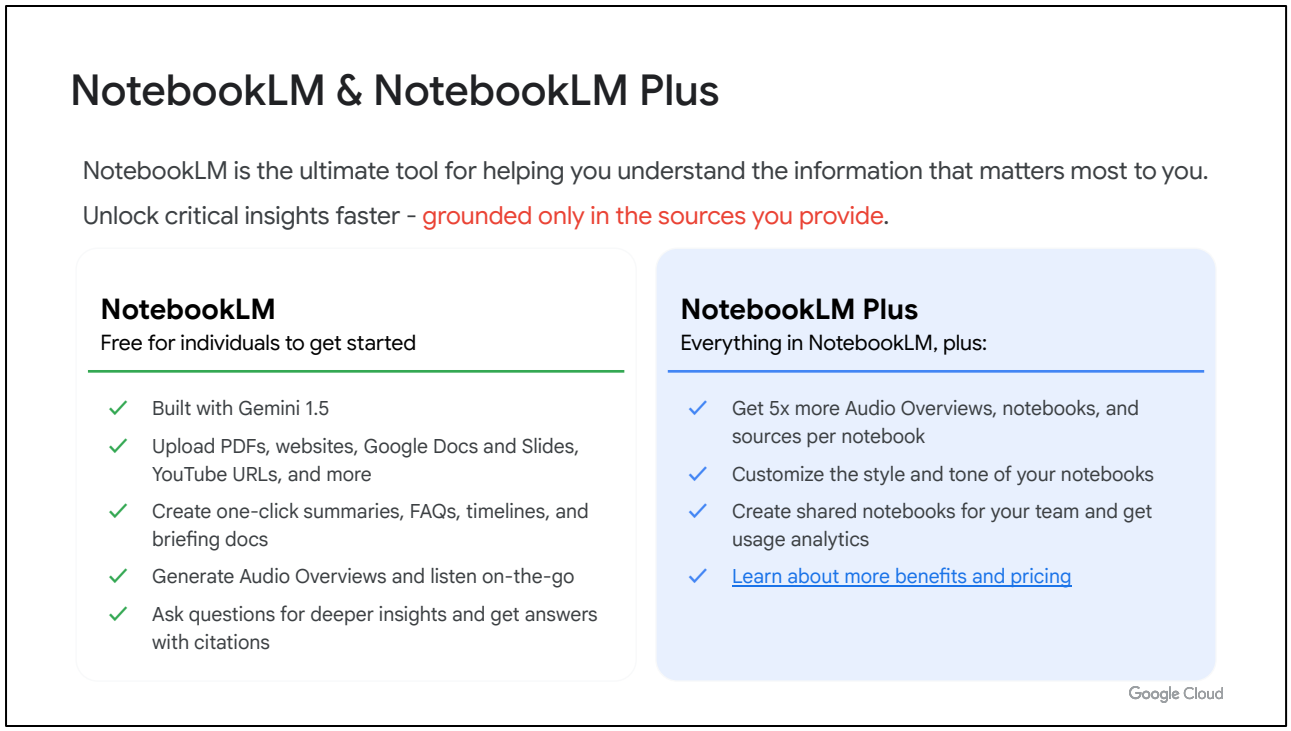
Gemini Enterprise example use-case



Google Cloud

SHOW DEMOS FROM:
https://cloud.google.com/gemini-enterprise?hl=en (for example, Marketing and HR)

This use case demonstrates Agentspace's ability to handle product innovation. This particular agent synthesizes a bunch of information from multiple sources, creating content options, and sharing a project briefing across a large team.



Let's switch from Gemini app (and various add-ons) to another tool, which we've seen last time. What we said last week was: "NotebookLM is a great tool for learning based on specific set of resources" - and because this tool is "grounded" to those resources, it should not hallucinate.

For the exam, you need to differentiate between standard version of this tool and so-called "NotebookLM Plus", which basically increases daily limits, but it also allows you to share your notebooks with your teams - and I've personally got a question on the exam about "collaboration", where NotebookLM Plus was the proper choice.

Gen AI Risks

There are five top security risks which should be addressed for AI. These are:

1) Prompt Injection (getting the model to execute malicious instructions)

EXAMPLES:

- trick the AI into ignoring its original instructions and constraints and trying to convince car-sales chatbot to sell a car for 1$... maybe not harmful to a human being, but definitely to your business if you're the dealer

2) Data Exposure (a model revealing private data)

EXAMPLE: by asking a specific question to a company's custom chatbot, users might persuade it to reveal the confidential documents and data it was trained on

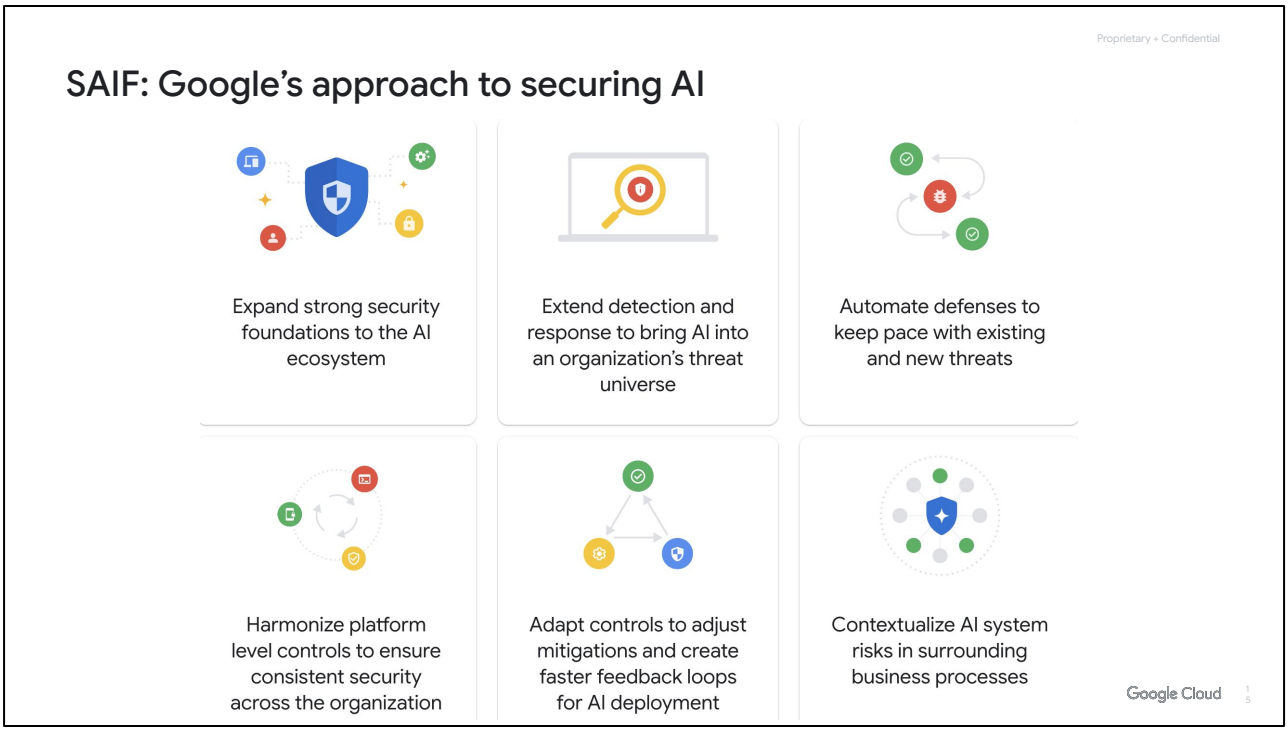
3) Model Theft (a model being stolen by an attacker)

4) Data Poisoning (altering a dataset to change model behavior);

Example: there was an AI chatbot on one of the social platforms designed to learn from conversations. Bad actors intentionally fed it racist and inflammatory content. Within a day, the chatbot's training

data was so poisoned that it began generating offensive and hateful information
and

5) Model Integrity Compromise (tampering with the model to change its behavior toward a harmful outcome)



The Secure AI Framework (SAIF) is Google's blueprint released in June 2023 to help navigate the complex security challenges of the AI landscape, especially with the rise of generative AI. It provides clear guidance and best practices for building security into every stage of the AI lifecycle, from development to deployment.

SHOW: https://safety.google/cybersecurity-advancements/saif/ -> Download PDF

It offers a structured approach to address these challenges head-on.

By adhering to the six pillars of SAIF, organizations can establish a robust security posture for their AI initiatives. This means they can confidently innovate with AI, knowing their systems are secure, their data is protected, and their AI development aligns with responsible practices.

Google's approach to securing AI is built on the Secure AI Framework (SAIF). SAIF provides security standards and best practices by focusing on four distinct layers:

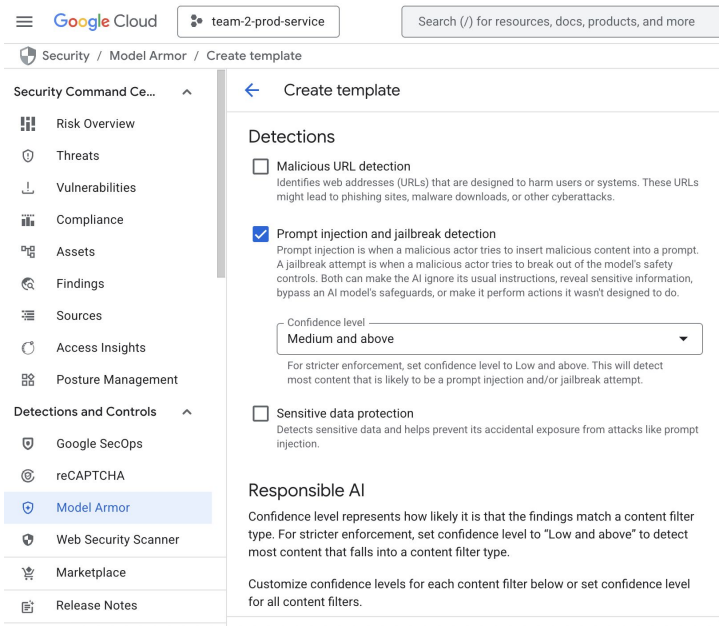
capabilities);

MODEL (protecting the model itself and its development process);

APPLICATION (protecting Gen AI applications, whether customer-, Google-, or vendor-managed); and

INFRASTRUCTURE (protecting underlying platforms and infrastructure, ensuring compliance and control)

Model Armor



Model Armor is one of the products that helps you **implement the principles** of the Secure AI Framework (SAIF).

**** Demo in SCC if possible. If not, explain the screenshot

It's tool within Google Cloud for detecting and mitigating risks. It allows setting detection options for threats like Prompt injection and jailbreak detection, and sensitive data protection. Users can customize confidence levels (e.g., set to "Medium and above" for stricter enforcement) for prompt injection and content filters related to Responsible AI

A screening tool for LLMs.

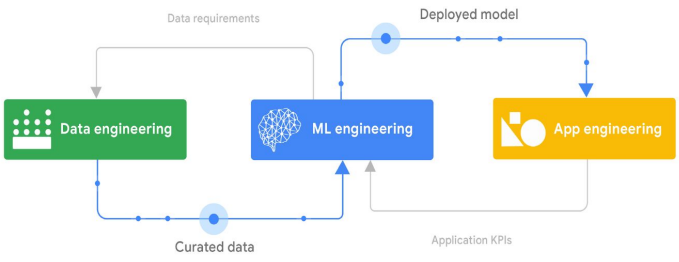
Basic functionality: scans the prompts (for any violations, malutious URLs etc) and model responses (ie. For sensitive data leakage) for malicious content etc.

Model Armor protects from 4 out of 10 AI-related OWASP threats:

Malicious files and unsafe URLs
Prompt injection and jailbreaks -> ie. Trying to convince car-sales chatbot to sell a car for 1$...
Sensitive data
Offensive material

MLOps

MLOps is an ML engineering culture and practice that aims at unifying a set of **standardized processes and capabilities** for building, deploying, and operationalizing ML systems **rapidly and reliably**.



Source: Practitioners guide to MLOps White Paper

Google Cloud

"MLOps" - definition.

Think of it as DevOps for Machine Learning.

It's iterative!

Generative AI projects follow an iterative path:

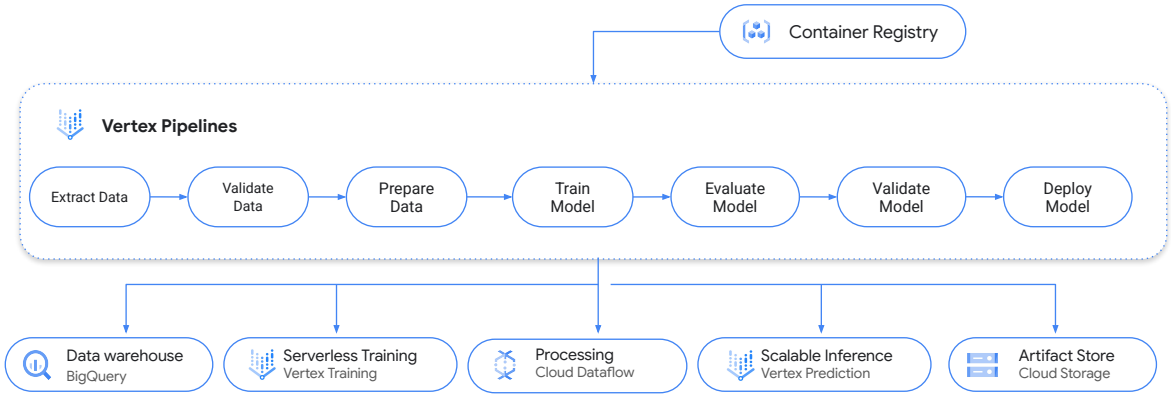
- **Define business use cases:** What problem are you trying to solve?
- **Data exploration:** Understand your data.
- **Select algorithm:** Choose the right model for the job.
- **Data Pipeline and Feature engineering:** Get your data ready and create useful features.
- **Build ML model:** Train your AI.
- **Evaluate:** See how well it performs.
- **Present results:** Share your findings.
- **Plan for Deployment:** How will it go live?
- **Operationalize Model:** Get it running in production.
- **Monitor Model:** Keep an eye on its performance.
- **Iterate on approach:** Continuously improve!

Full ML Lifecycle Mapping

MLOps Stage	Key Google Cloud Services
1. Gather Data	Pub/Sub, Cloud Storage, Cloud SQL, Cloud Spanner
2. Prepare Data	BigQuery, BigQuery universal catalog
3. Train Model	Vertex AI platform
4. Deploy & Predict	Vertex AI
5. Manage Model	Vertex AI Pipelines, Vertex AI Feature Store, Vertex AI Model Garden

Let’s have a look at a summary of those steps we covered from GCP service perspective.
First, you can argue that the Vertex AI platform (with various tools it provides can handle everything around models - from training new ones, updating their versions, all the way to deploying and using them.
When it comes to handling data, we really see quite a typical “Big Data” tools from Google, with Bigquery at the center, since it’s Google’s Data Warehouse, typically used whenever your use-case is mostly about analytics and “reading” data rather than performing transactions to create new data and update existing entries.

Vertex AI Pipelines: scalable and cost effective



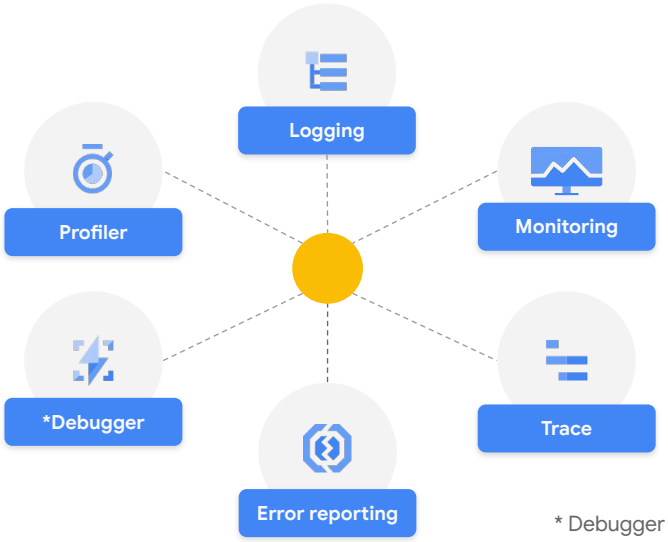
Vertex AI provides a suite of products to help you at each stage of your ML workflow, from gathering data, to feature engineering, to building models, and finally deploying and monitoring those models

Cloud Operations

Cloud Operations Suite (CloudWatch)

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

Cloud Operations Services

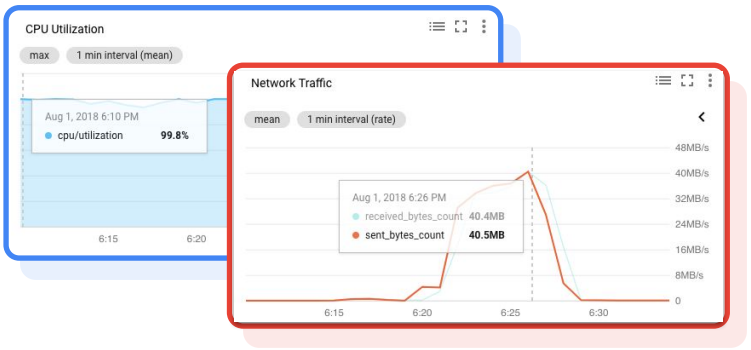


* Debugger Discontinued May 2023

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

Cloud Monitoring (CloudWatch) dashboards can visualize utilization and network traffic

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



Cloud Monitoring

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

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

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

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

Custom metrics

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

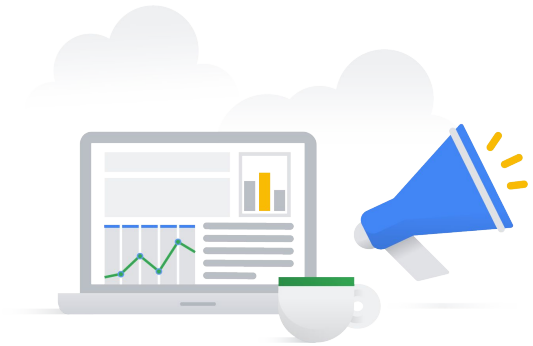
Custom metrics
https://cloud.google.com/monitoring/custom-metrics

OpenCensus: https://opencensus.io/

Alerting Policies

Create notifications in response to a policy that exceeds some condition

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



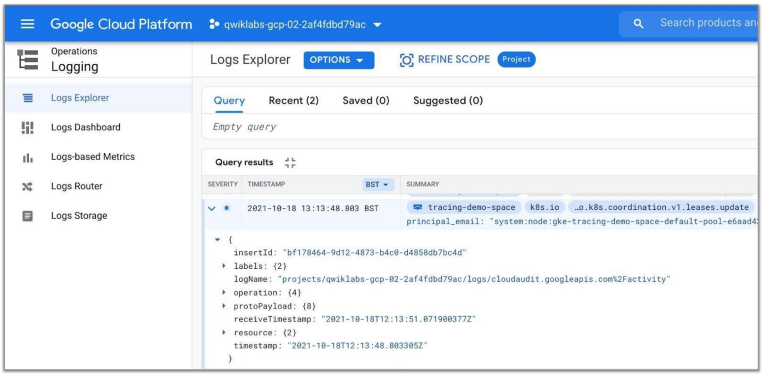
An alerting policy allows you to respond to anomalies in your system. For example, if your system starts to get an unusually high number of requests, you might want to be notified of a potential denial of service attack. Or, if a VM is not working, you can have the system recreate it.

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

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

Cloud Logging (CloudWatch Logs) collects logs from admin, system and application activity

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

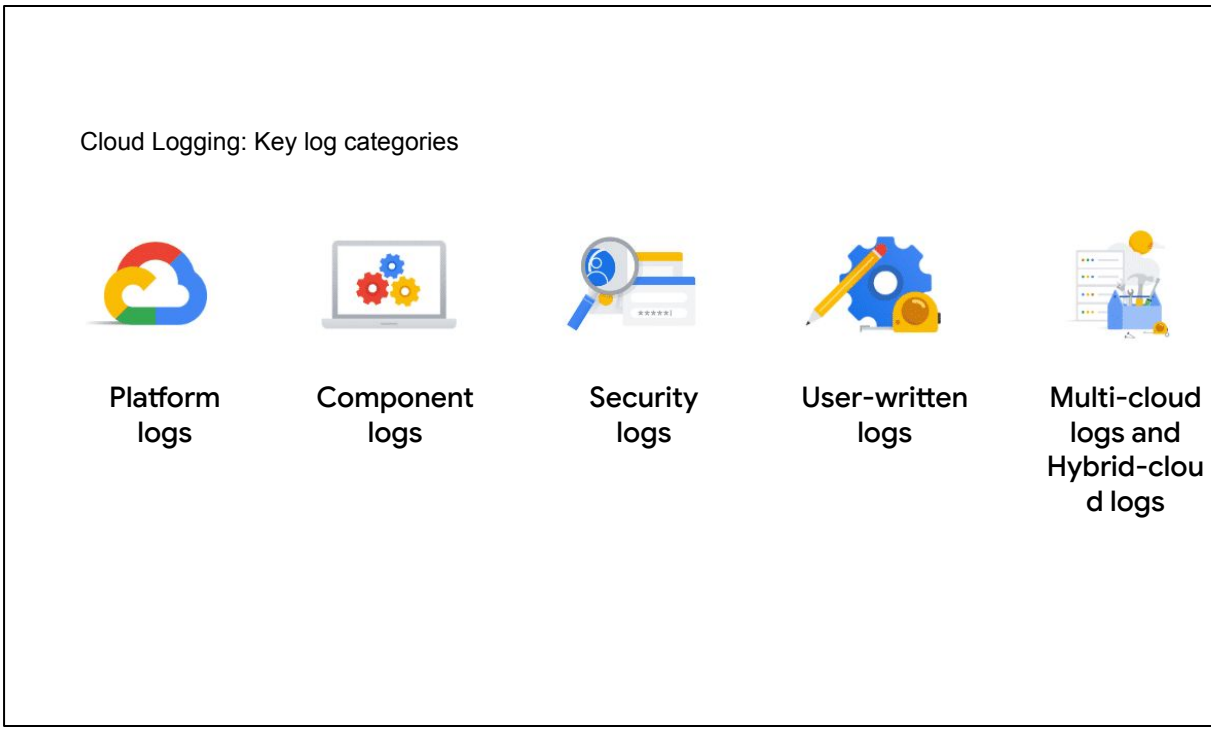


Cloud Logging pricing for Cloud Admins: How to approach it & save cost

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

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

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



The Google Cloud platform logs visible to you in Cloud Logging vary, depending on which Google Cloud resources you're using in your Google Cloud project or organization.

Let's explore the key log categories.

Platform logs are logs written by your Google Cloud services.

These logs can help you debug and troubleshoot issues, and help you better understand the Google Cloud services you're using.

For example, VPC Flow Logs record a sample of network flows sent from and received by VM instances.

Component logs are similar to platform logs, but they are generated by Google-provided software components that run on your systems.

For example, GKE provides software components that users can run on their own VM or in their own data center. Logs are generated from the user's GKE instances and sent to a user's Cloud project. GKE uses the logs or their metadata to provide user support.

Security logs help you answer "who did what, where, and when."

- Cloud Audit Logs provide information about administrative activities and

- accesses within your Google Cloud resources.
- Access Transparency provides you with logs of actions taken by Google staff when accessing your Google Cloud content.

User-written logs are logs written by custom applications and services.

Typically, these logs are written to Cloud Logging by using one of the following methods:

- Ops Agent
- Cloud Logging API
- Cloud Logging client libraries

Finally, there Multi-cloud logs and Hybrid-cloud logs.

These refer to logs from other cloud providers like Microsoft Azure and logs from on-premises infrastructure.

Cloud Logging: Key log categories (1/2)

Platform logs

Logs written by your Google Cloud services

Debug and troubleshoot issues

Better understand Google Cloud services

Component logs

Similar to platform logs, but are generated by Google-provided software components that run on your systems

Security logs

Help you answer "who did what, where, and when"

Information about administrative activities and accesses

Logs of actions

The Google Cloud platform logs visible to you in Cloud Logging vary, depending on which Google Cloud resources you're using in your Google Cloud project or organization.

Let's explore the key log categories.

Platform logs are logs written by your Google Cloud services.

These logs can help you debug and troubleshoot issues, and help you better understand the Google Cloud services you're using.

For example, VPC Flow Logs record a sample of network flows sent from and received by VM instances.

Component logs are similar to platform logs, but they are generated by Google-provided software components that run on your systems.

For example, GKE provides software components that users can run on their own VM or in their own data center. Logs are generated from the user's GKE instances and sent to a user's Cloud project. GKE uses the logs or their metadata to provide user support.

Security logs help you answer "who did what, where, and when."

- Cloud Audit Logs provide information about administrative activities and

- accesses within your Google Cloud resources.
- Access Transparency provides you with logs of actions taken by Google staff when accessing your Google Cloud content.

User-written logs are logs written by custom applications and services.

Typically, these logs are written to Cloud Logging by using one of the following methods:

- Ops Agent
- Cloud Logging API
- Cloud Logging client libraries

Finally, there Multi-cloud logs and Hybrid-cloud logs.

These refer to logs from other cloud providers like Microsoft Azure and logs from on-premises infrastructure.

Cloud Logging: Key log categories (2/2)

User-written logs

Logs written by custom applications and services

- Ops Agent
- Cloud Logging API
- Cloud Logging client libraries

Multi- and Hybrid-cloud logs

Logs from other cloud providers like Microsoft Azure and logs from on-premises infrastructure

The Google Cloud platform logs visible to you in Cloud Logging vary, depending on which Google Cloud resources you're using in your Google Cloud project or organization.

Let's explore the key log categories.

Platform logs are logs written by your Google Cloud services.

These logs can help you debug and troubleshoot issues, and help you better understand the Google Cloud services you're using.

For example, VPC Flow Logs record a sample of network flows sent from and received by VM instances.

Component logs are similar to platform logs, but they are generated by Google-provided software components that run on your systems.

For example, GKE provides software components that users can run on their own VM or in their own data center. Logs are generated from the user's GKE instances and sent to a user's Cloud project. GKE uses the logs or their metadata to provide user support.

Security logs help you answer "who did what, where, and when."

- Cloud Audit Logs provide information about administrative activities and

- accesses within your Google Cloud resources.
- Access Transparency provides you with logs of actions taken by Google staff when accessing your Google Cloud content.

User-written logs are logs written by custom applications and services.

Typically, these logs are written to Cloud Logging by using one of the following methods:

- Ops Agent
- Cloud Logging API
- Cloud Logging client libraries

Finally, there Multi-cloud logs and Hybrid-cloud logs.

These refer to logs from other cloud providers like Microsoft Azure and logs from on-premises infrastructure.

Log Sinks (Export)

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

- To export specific logs write a filter that selects them, and choose a destination in Cloud Storage (S3), BigQuery, Pub/Sub, Cloud Logging, OpenSearch, or Splunk
  - The filter and destination are held in an object called a sink (written in a Lambda function)
    - Can be created at the organization, folder, project (account), and billing account level
- A sink can only export logs that belong to its parent resource, e.g. project/folder.
  - When a log comes in that matches a filter, a copy of the log is written to the export destination.

Sinks

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

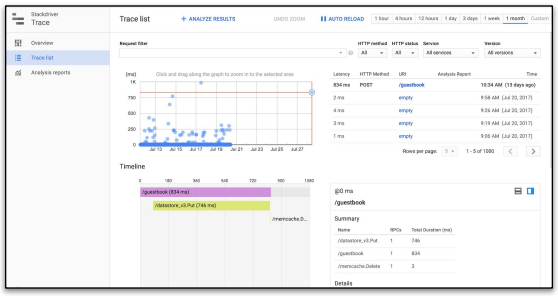
Scenarios for exporting Cloud Logging data - Splunk
https://cloud.google.com/architecture/exporting-stackdriver-logging-for-splunk

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

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

Cloud Trace (X-Ray)

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



Viewing trace details

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

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

Cloud Profiler (CodeGuru Profiler)

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

Cloud Profiler

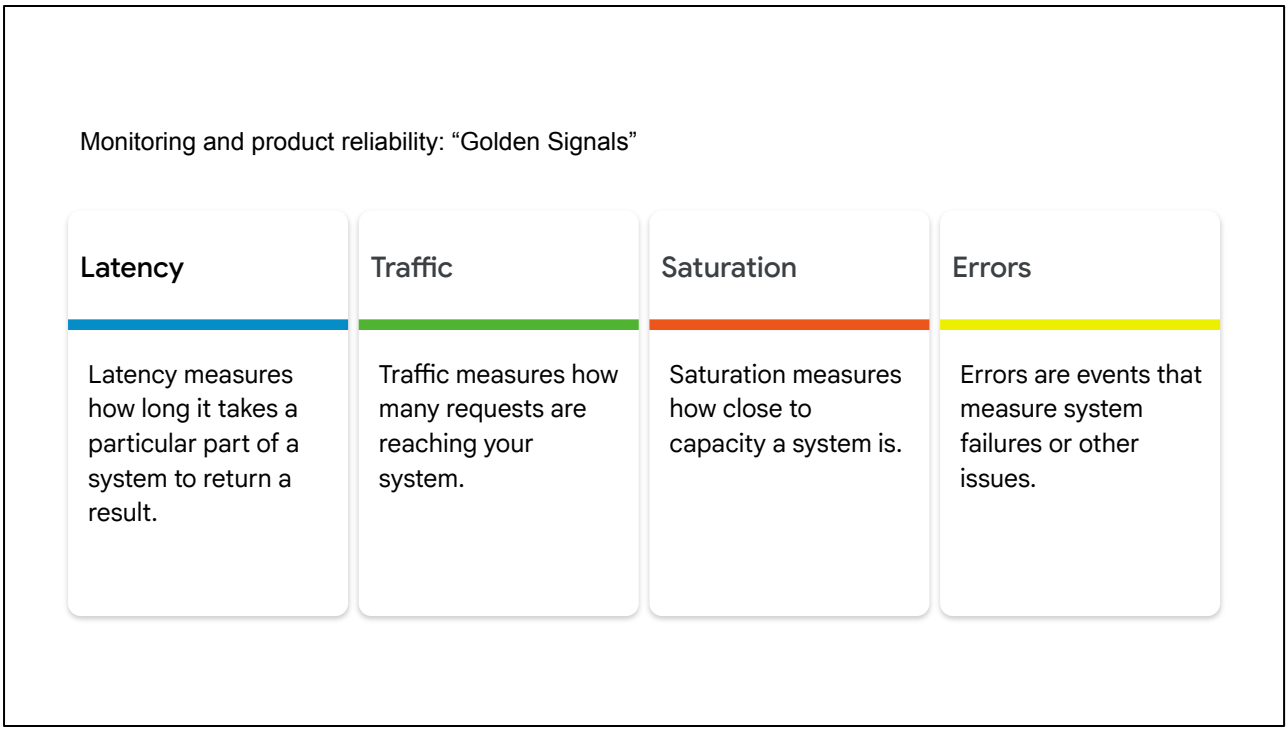
Cloud Profiler
https://cloud.google.com/profiler/docs

Cloud Error Reporting

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

Cloud Error Reporting

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



In Google's *Site Reliability Engineering* book, monitoring is defined as "collection, processing, aggregating, and displaying real-time quantitative data about a system, such as query counts and types, error counts and types, processing times, and server lifetimes."

There are “Four Golden Signals” that measure a system’s performance and reliability.

Let's explore these measures.

First is latency.

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

Sample latency metrics include:

- Page load latency
- Number of requests waiting for a thread
- Query duration
- Service response time
- Transaction duration
- Time to first response
- Time to complete data return

Next, traffic measures how many requests are reaching your system.

Sample traffic metrics include the number of:

- HTTP requests per second
- requests for static vs. dynamic content
- concurrent sessions
- transactions per second
- retrievals per second
- active requests
- write ops
- read ops, and
- active connections

Traffic metrics also include Network Input/Output (or I/O).

Saturation measures how close to capacity a system is.

Capacity is often a subjective measure that depends on the underlying service or application.

Sample capacity metrics include:

- The percentage of:
  - memory utilization
  - thread pool utilization
  - cache utilization
  - disk utilization
  - CPU utilization

- Other metrics include:
  - Disk quota
  - Memory quota
  - The number of available connections, and
  - The number of users on the system

Finally, errors are events that measure system failures or other issues.

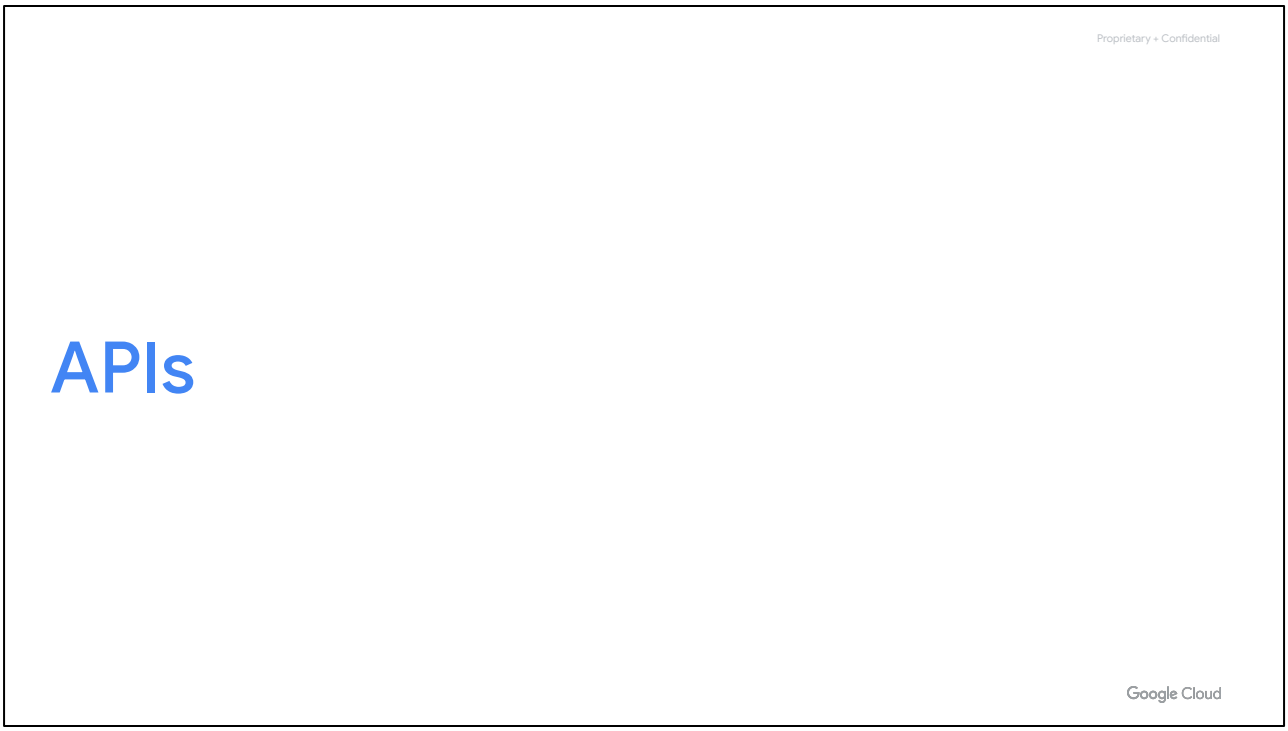
Errors are often raised when a flaw, failure, or fault in a computer program or system causes it to produce incorrect or unexpected results, or behave in unintended ways.

Sample error metrics include the number of:

- 400/500 HTTP codes
- failed requests
- exceptions
- stack traces
- dropped connections

Other metrics include:

- Wrong answers or incorrect content
- Servers that fail liveness checks



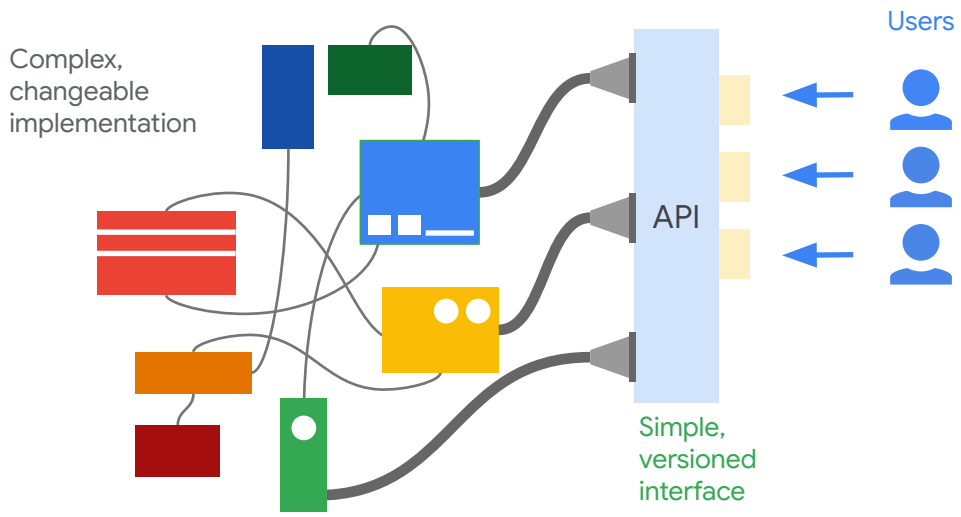
Source: Architecting with Compute Engine Slides

Let’s move our attention from hybrid connectivity to sharing VPC networks.

In the simplest cloud environment, a single project might have one VPC network, spanning many regions, with VM instances hosting very large and complicated applications. However, many organizations commonly deploy multiple, isolated projects with multiple VPC networks and subnets.

In this lesson, we are going to cover two configurations for sharing VPC networks across Google Cloud projects. First, we will go over shared VPC, which allows you to share a network across several projects in your Google Cloud organization. Then, we will go over VPC Network Peering, which allows you to configure private communication across projects in the same or different organizations.

APIs hide the details and enforce contracts

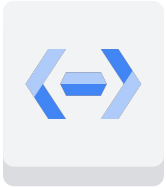


Managing APIs: Cloud Endpoints / Apigee (API Gateway)

Both provide tools for:

- User authentication
- Monitoring
- Securing APIs
- Etc.

Both support OpenAPI and gRPC



Cloud Endpoints
(API Gateway)

Google Cloud Only
2 million/mth free

Apigee API
Platform
(API Gateway)

Hybrid Deployments
Monetization feature

Apigee
https://docs.apigee.com/

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

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

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

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

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

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

Cloud Endpoints helps you create and maintain APIs

- Distributed API management through an API console.
- Expose your API using a RESTful interface.
- Control access and validate calls with JSON Web Tokens and Google API keys (AWS Signature Version 4 or Lambda Authorizer).
  - Identify web, mobile users with Auth0 and Firebase (Cognito) Authentication.
- Generate client libraries.

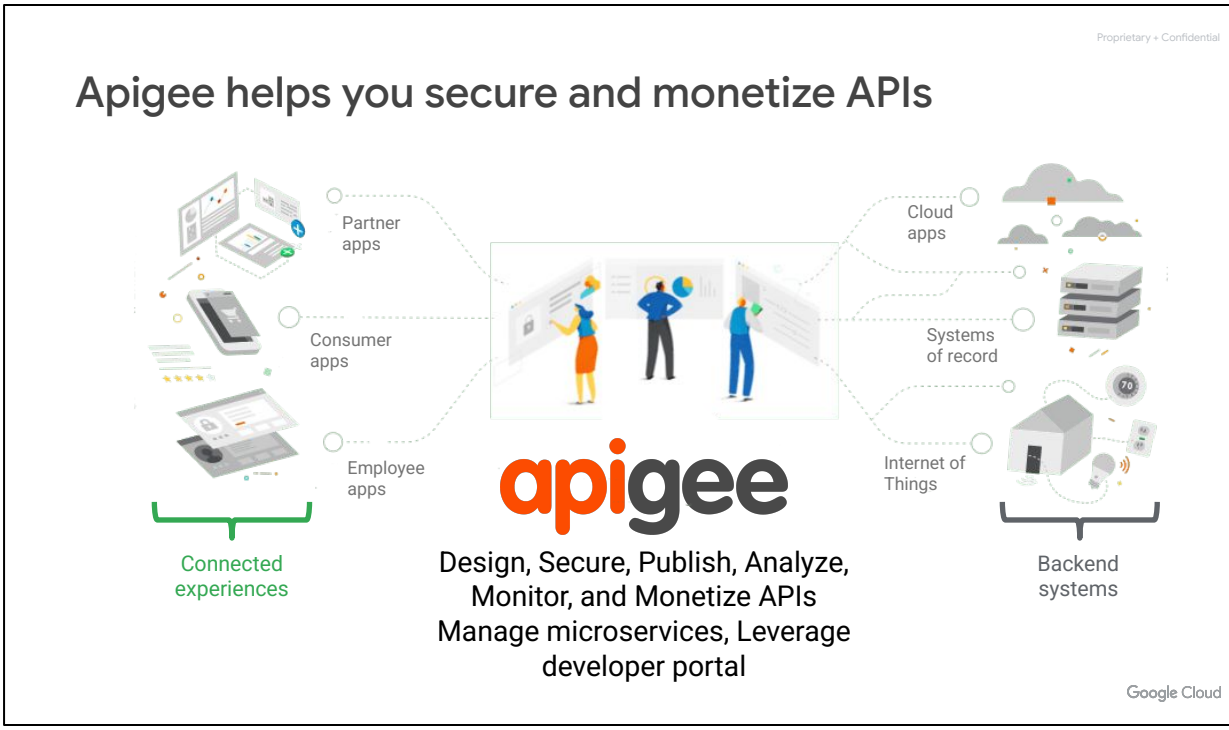
Source: Demo Template

Cloud Endpoints is a distributed API management system. It provides an API console, hosting, logging, monitoring, and other features to help you create, share, maintain, and secure your APIs. You can use Cloud Endpoints with any APIs that support the OpenAPI Specification, formerly known as the Swagger spec.

Cloud Endpoints uses the distributed Extensible Service Proxy to provide low latency and high performance for serving even the most demanding APIs. Extensible Service Proxy is a service proxy based on NGINX. It runs in its own Docker container for better isolation and scalability. The proxy is containerized and distributed in the Container Registry and Docker registry, and can be used with App Engine, Google Kubernetes Engine, Compute Engine or Kubernetes.

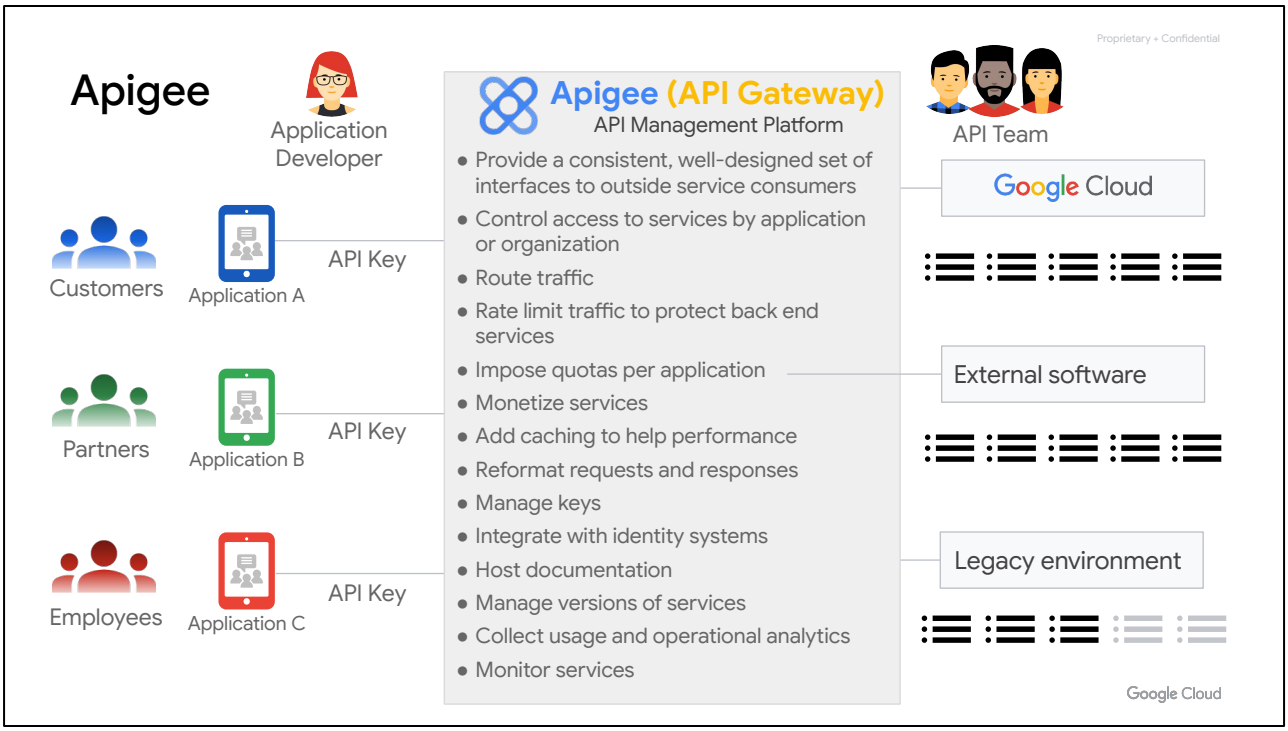
Cloud Endpoints features
User authentication: JSON Web Token validation and a streamlined developer experience for Firebase Auth, Google Auth and Auth0.
Automated deployment: With App Engine, the proxy is deployed automatically with your application. On Google Kubernetes Engine or Compute Engine, use Google's containerized ESP for simple deployment.
Logging and monitoring: Monitor traffic, error rates and latency, and review logs

in Cloud Logging. Use Cloud Trace to dive into performance and BigQuery for analysis.
API keys: Generate API keys in the Cloud Console and validate on every API call. Share your API with other developers to allow them to generate their own keys.
Easy integration: Get started quickly by using one of Google's Cloud Endpoints Frameworks or by simply adding an Open API specification to your deployment.



Source: Demo Template

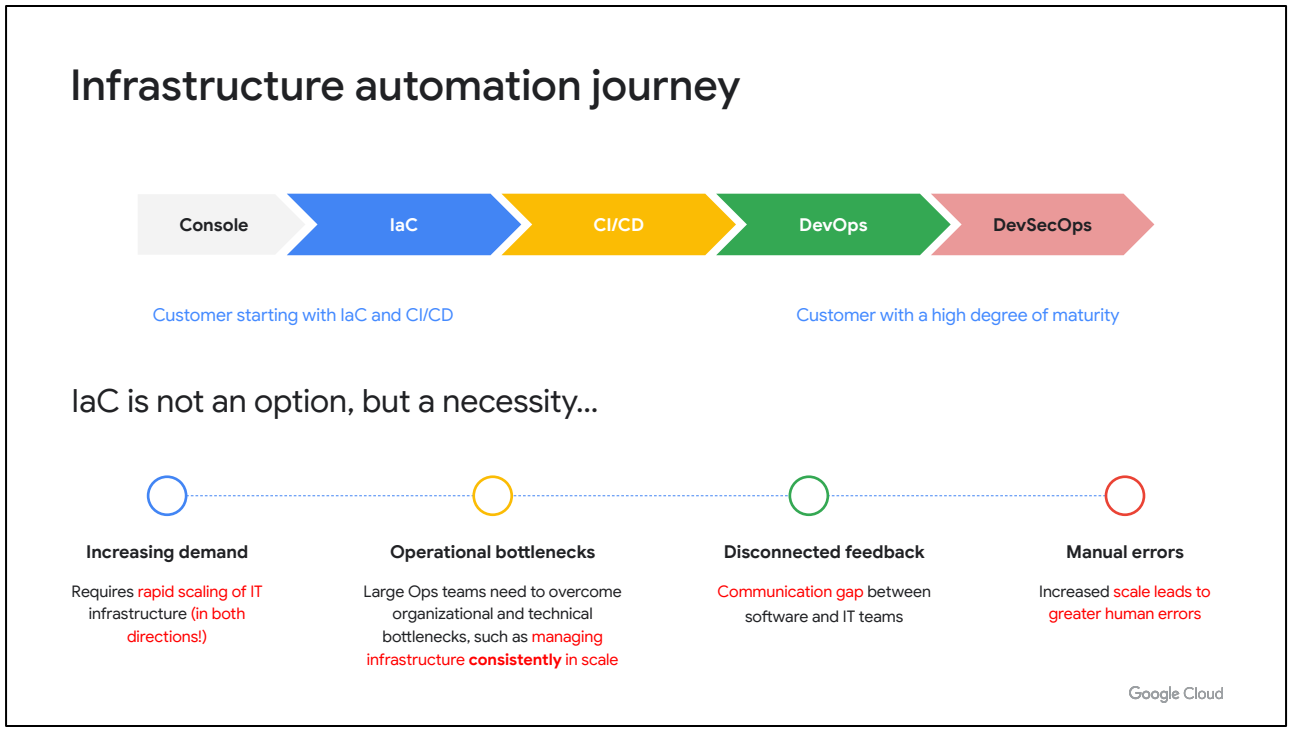
Apigee Edge is also a platform for developing and managing API proxies. It has a different orientation, though: it has a focus on business problems like rate limiting, quotas, and analytics. Many users of Apigee Edge are providing a software service to other companies, and those features come in handy. Because the backend services for Apigee Edge need not be in Google Cloud, engineers also often use it when they are working to take a legacy application apart. Instead of replacing a monolithic application in one risky move, they can instead use Apigee Edge to peel off its services one by one, standing up microservices to implement each in turn, until the legacy application can finally be retired.


Apigee
Application Developer
Customers
Application A
API Key
Partners
Application B
API Key
Employees
Application C
API Key
Apigee (API Gateway)
API Management Platform
Provide a consistent, well-designed set of interfaces to outside service consumers
Control access to services by application or organization
Route traffic
Rate limit traffic to protect back end services
Impose quotas per application
Monetize services
Add caching to help performance
Reformat requests and responses
Manage keys
Integrate with identity systems
Host documentation
Manage versions of services
Collect usage and operational analytics
Monitor services
API Team
Google Cloud
External software
Legacy environment
Proprietary + Confidential
Google Cloud

Managing Implementation

Google Cloud

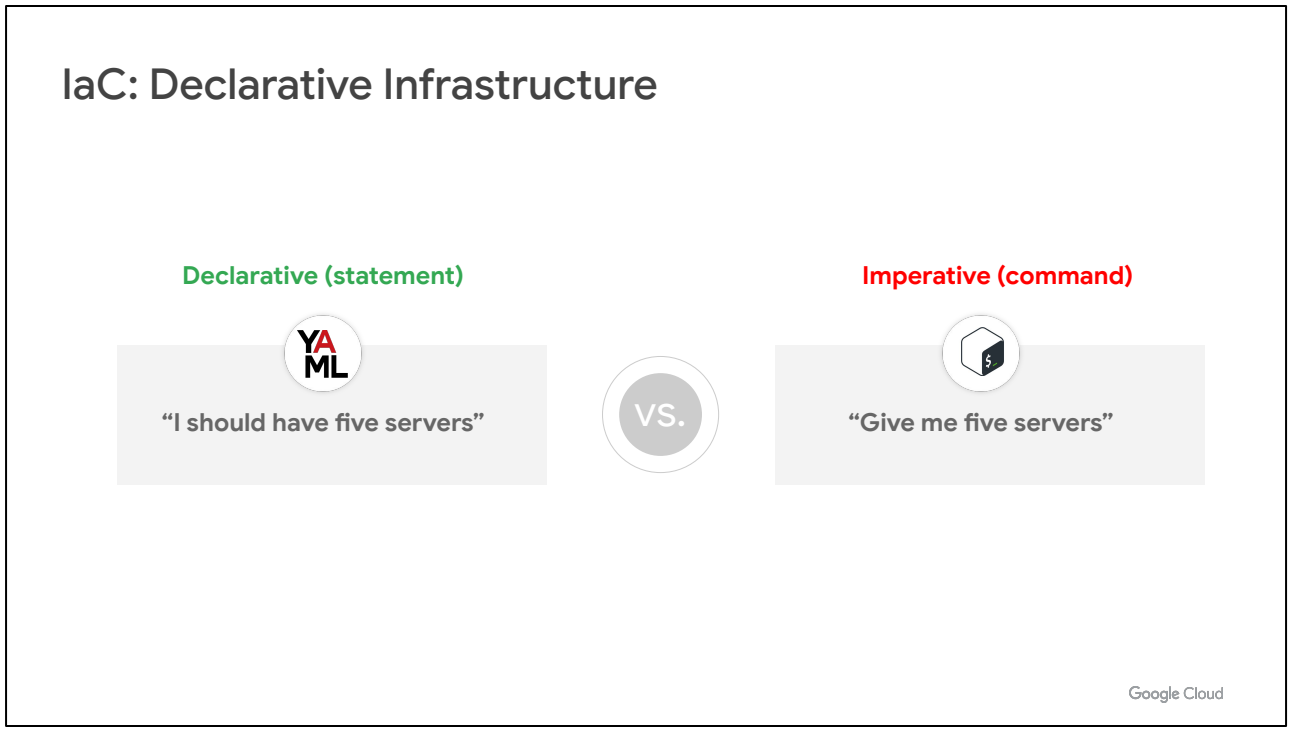
Let’s begin by considering the role of a Professional Cloud Architect in implementation, operations, and reliability at Cymbal Direct.



– typical journey, which evolves as cloud customers get more mature
– important that some automation concepts are used from the very beginning (LZ deployment), as it might be very difficult to switch from manual management to automation later on (talk about an example customer)

- DevSecOps means shifting the security review process "left" or earlier in the software development lifecycle

- IaC is not an option but a necessity due to:
- **Increasing demand.** High business demand has been necessitating the **rapid scaling of IT infrastructure** backbones across industries.
- **Operational bottlenecks.** With the rapid scaling of IT Infrastructure, **Ops teams** need to **overcome** new **organizational and technical bottlenecks**, such as managing infrastructure consistently in scale
- **Disconnected feedback loops.** The need to close the communication gap **between software and IT teams** is becoming **imperative for successful deployments.**
- **Manual errors.** Increased quantity and scale has led to greater human error with the potential for significant impacts.



– state management is an important part of IaC since whatever tool we’re using, it needs to know what’s the current state to plan the necessary changes.

Declarative IaC focuses on the desired statement ("I should have five servers"). Imperative IaC (like shell commands) focuses on the execution command ("Give me five servers").

With infrastructure as code, we prefer **declarative** infrastructure where you state how the infrastructure should be and not the commands to take.

GCP services emulators

for local development, testing and validation

- Spanner:
  - https://cloud.google.com/spanner/docs/emulator
  - locally-running, emulated instance of Cloud Spanner to enable local development and testing.
  - https://github.com/GoogleCloudPlatform/cloud-spanner-emulator
- Pub/Sub:
  - https://cloud.google.com/pubsub/docs/emulator
- Bigtable:
  - https://cloud.google.com/bigtable/docs/emulator
- Firestore:
  - https://cloud.google.com/firestore/docs/emulator
- Cloud Run:
  - https://cloud.google.com/run/docs/testing/local

Google Cloud

GCP provides service emulators for local development, testing, and validation. Emulators are available for key services such as Spanner, Pub/Sub, Bigtable, Firestore, and Cloud Run. These allow local development and testing without incurring cloud costs

Cloud Shell (Part of Cloud Console)

- A temporary Debian based, Compute Engine virtual machine instance in a web browser
- Built-in code editor
- 5 GB of persistent disk storage
- Pre-installed Google Cloud SDK and other tools
- Web preview functionality
- Built-in authorization for access to Google Cloud Console projects and resources

Google Cloud

Cloud Shell is a temporary Debian-based, Compute Engine virtual machine instance accessible via a web browser. It features a built-in code editor, 5 GB of persistent disk storage, a pre-installed Google Cloud SDK and other tools, web preview functionality, and built-in authorization for accessing Google Cloud Console projects and resources

Cloud Shell Editor

Cloud Shell also provides a VS Code-like web IDE.

The IDE runs on the Cloud Shell VM, and serves a web interface on port 970. The Cloud Shell UI iframes this web interface and provides additional UI and features around it.

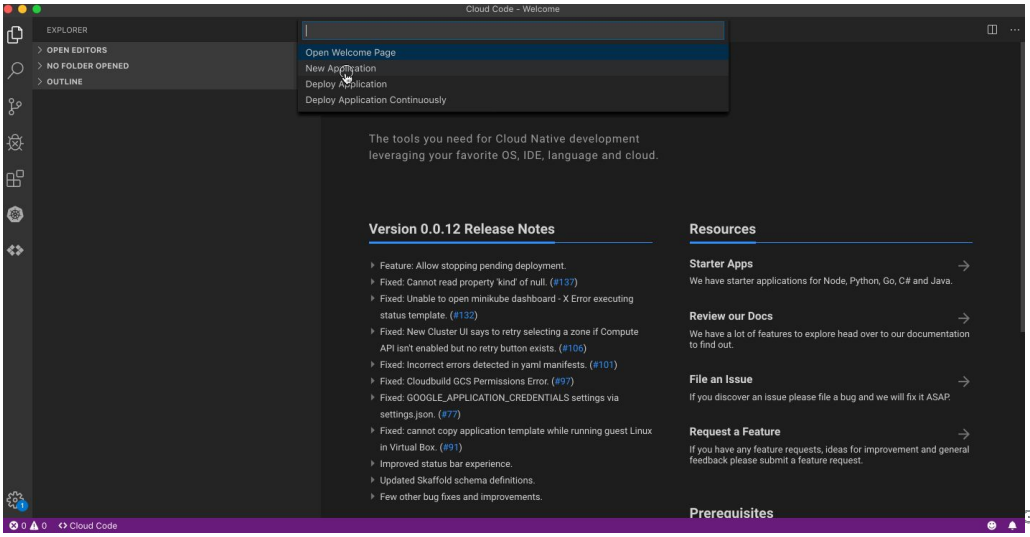
We maintain a custom Cloud Shell extension that runs on the IDE and manages communication with the rest of the Cloud Shell UI. This extension allows for interactions such as managing the app lifecycle (loading, reconnect) in the Cloud Shell UI, or allowing Neos to open specific files in the Editor or spotlight UI elements.

Google Cloud

The Cloud Shell Editor is a VS Code-like web IDE that runs on the Cloud Shell VM. It serves a web interface on port 970 and provides additional UI features and extensions for managing the app lifecycle

– DEMO?

Cloud Code



https://github.com/GoogleCloudPlatform/cloud-code-samples/blob/v1/README.md

Cloud Code is basically a range of extensions that should increase the productivity for developers. It consists of extensions for IDEs: Visual Studio Code and IntelliJ Platform - to surface GCP integrations that feel like are working with local code:

a. – SHOW example for IntelliJ: https://plugins.jetbrains.com/plugin/8079-gemini-code-assist-cloud-code

Cloud Code extension is also added also added to Cloud Shell Editor:

- SHOW Editor in Cloud Console

These extensions make it easy to **create and deploy applications to multiple Google Cloud services** such as GKE or Cloud Run.

It also enable a **fast code/build/test cycle** with hot reload, possible by built-in integrations with Minikube and Skaffold

- To get some hands on practice, go through this codelab:

(SHOW, go to task 1, explain Minikube etc)

Cymbal Retail

case study

https://services.google.com/fh/files/misc/v6.1_pca_cymbal_retail_case_study_english.pdf

Helicopter Racing League (HRL) is a global sports league for competitive helicopter racing. Each year HRL holds the world championship and several regional league competitions where teams compete to earn a spot in the world championship. HRL offers a paid service to stream the races all over the world with live telemetry and predictions throughout each race.

Proposed Technical Solution

- **Catalog and Content Enrichment**:
  - Vertex AI with Gemini Models to extract structured attributes, generate compelling product descriptions and enrich data.
  - Vertex AI Search for Commerce - service specifically designed for retail and e-commerce (power search, recommendations, catalog improvement) with HITL-enabled.
  - Cloud Vision API and Document AI (with Human-in-the-Loop… deprecated?) for extracting information from visual and document-based supplier data. Imagen on Vertex AI for image creation
  - Storage part: Cloud Storage for data hosting, and BigQuery for storing and normalizing the final attributes.
- **Conversational Commerce with Product Discovery**:
  - Conversational Commerce agent (with IVR enabled and Security Settings for data) connected to Vertex AI Search for Commerce service to use product catalog as its primary source of information. Have this agent deployed to website and mobile app by following developer's guide and / or using Dialogflow CX.
  - Potentially, implement CCAI to move away from legacy IVR and manual agent processes, reduce call center costs, and implement modern AI-powered conversational agents for product discovery and sales conversion
- **Technical Stack Modernization**:
  - Migrate on-prem Kubernetes to GKE (Autopilot) or Cloud Run where possible.
  - Migrate relational databases (MySQL, MS SQL) to managed services like Cloud SQL. BigQuery might be suitable for consolidating product catalog and customer data, overcoming data silos and supporting the Gen AI needs. Memorystore can manage Redis instances.
  - Utilize tools like Cloud Data Fusion or Cloud Composer for modern, efficient orchestration and transformation of data, replacing legacy ETL processes. Modernize 3rd party integrations using an API management platform (Apigee).
  - Modernize monitoring with Cloud Logging, Cloud Monitoring, Uptime Checks, Alerting Policies, Managed Service for Prometheus etc
  - Security and compliance: KMS, encrypted hybrid connectivity, SDP for sensitive data, VPC-SC perimeters around sensitive APIs

[Cymbal Retail case study] Diagnostic Question #1

Cymbal Retail's current environment has data silos across various databases (MySQL, Microsoft SQL Server, Redis, and MongoDB), which limits a unified view of the customer. They want to create a centralized data platform to get a holistic view of their customers and enable advanced analytics.

Which Google Cloud service would be the most appropriate foundation for this centralized data platform?

A. Cloud SQL

B. Bigtable

C. Firestore

D. BigQuery

Google Cloud

D

Answer: D

Explanation: BigQuery is a fully managed, serverless data warehouse that is designed for large-scale data analytics. It can ingest data from various sources, including the databases Cymbal Retail is currently using, and provides a unified platform for analysis. This directly addresses the problem of data silos. Cloud SQL is a managed relational database service, but it's not the best choice for a data warehouse. Bigtable is a NoSQL database suitable for large analytical and operational workloads, but BigQuery is more optimized for the kind of analytical queries needed for a unified customer view. Firestore is a NoSQL document database for mobile, web, and server development that is not a data warehousing solution.

[Cymbal Retail case study] Diagnostic Question #1

Cymbal Retail's current environment has data silos across various databases (MySQL, Microsoft SQL Server, Redis, and MongoDB), which limits a unified view of the customer. They want to create a centralized data platform to get a holistic view of their customers and enable advanced analytics.

Which Google Cloud service would be the most appropriate foundation for this centralized data platform?

A. Cloud SQL

B. Bigtable

C. Firestore

D. BigQuery

Google Cloud

D

Answer: D

Explanation: BigQuery is a fully managed, serverless data warehouse that is designed for large-scale data analytics. It can ingest data from various sources, including the databases Cymbal Retail is currently using, and provides a unified platform for analysis. This directly addresses the problem of data silos. Cloud SQL is a managed relational database service, but it's not the best choice for a data warehouse. Bigtable is a NoSQL database suitable for large analytical and operational workloads, but BigQuery is more optimized for the kind of analytical queries needed for a unified customer view. Firestore is a NoSQL document database for mobile, web, and server development that is not a data warehousing solution.

[Cymbal Retail case study] Diagnostic Question #2

Cymbal Retail's custom-built web application experiences significant traffic fluctuations. They want to ensure that the application can handle traffic spikes without manual intervention while minimizing costs. The application is containerized and runs on Kubernetes.

Which GKE feature should they leverage to achieve this?

A. Horizontal Pod Autoscaler (HPA)

B. Vertical Pod Autoscaler (VPA)

C. Cluster Autoscaler

D. A combination of HPA and Cluster Autoscaler

Google Cloud

Answer: D

Explanation: A combination of the **Horizontal Pod Autoscaler (HPA)** and the **Cluster Autoscaler** will provide the best solution. The **HPA** automatically scales the number of pods in a deployment based on CPU utilization or other custom metrics, which is perfect for handling traffic spikes. The **Cluster Autoscaler** automatically resizes the GKE cluster by adding or removing nodes based on the resource demands of the pods. Using both together ensures that there are enough pods to handle the load and enough nodes in the cluster to run those pods, all without manual intervention and in a cost-effective manner.

[Cymbal Retail case study] Diagnostic Question #2

Cymbal Retail's custom-built web application experiences significant traffic fluctuations. They want to ensure that the application can handle traffic spikes without manual intervention while minimizing costs. The application is containerized and runs on Kubernetes.

Which GKE feature should they leverage to achieve this?

A. Horizontal Pod Autoscaler (HPA)

B. Vertical Pod Autoscaler (VPA)

C. Cluster Autoscaler

D. A combination of HPA and Cluster Autoscaler

Google Cloud

Answer: D

Explanation: A combination of the **Horizontal Pod Autoscaler (HPA)** and the **Cluster Autoscaler** will provide the best solution. The **HPA** automatically scales the number of pods in a deployment based on CPU utilization or other custom metrics, which is perfect for handling traffic spikes. The **Cluster Autoscaler** automatically resizes the GKE cluster by adding or removing nodes based on the resource demands of the pods. Using both together ensures that there are enough pods to handle the load and enough nodes in the cluster to run those pods, all without manual intervention and in a cost-effective manner.

[Cymbal Retail case study] Diagnostic Question #3

Cymbal Retail wants to implement a solution that can automatically generate high-quality images of their products in different settings and styles for their e-commerce website.

Which Vertex AI service and model type would be most appropriate for this task?

A. Vertex AI Language with a text generation model.

B. Vertex AI with a generative adversarial network (GAN) model.

C. Vertex AI Vision with an image generation model (e.g., Imagen).

D. Vertex AI Vision with a classification model.

Google Cloud

C

Vertex AI Vision, with a powerful image generation model like Imagen, is designed for creating high-quality, realistic images from text prompts, making it perfect for this use case.

[Cymbal Retail case study] Diagnostic Question #3

Cymbal Retail wants to implement a solution that can automatically generate high-quality images of their products in different settings and styles for their e-commerce website.

Which Vertex AI service and model type would be most appropriate for this task?

A. Vertex AI Language with a text generation model.

B. Vertex AI with a generative adversarial network (GAN) model.

C. Vertex AI Vision with an image generation model (e.g., Imagen).

D. Vertex AI Vision with a classification model.

Google Cloud

C

Vertex AI Vision, with a powerful image generation model like Imagen, is designed for creating high-quality, realistic images from text prompts, making it perfect for this use case.

[Cymbal Retail case study] Diagnostic Question #4

Cymbal Retail wants to use Generative AI to provide their customer service agents with real-time assistance. The system should listen to customer calls, transcribe them in real-time, and suggest relevant answers and solutions from a knowledge base.

Which combination of Google Cloud services would be best for this solution?

A. Speech-to-Text API, Vertex AI Language, and Cloud Storage.

B. Contact Center AI (CCAI) Platform, which integrates Speech-to-Text, Dialogflow, and Agent Assist.

C. Dialogflow CX, Speech-to-Text API, and BigQuery

D. Cloud Pub/Sub, Cloud Functions, and the Speech-to-Text API.

Google Cloud

B

The CCAI Platform is a purpose-built solution for this exact use case, providing a tightly integrated set of services for contact center automation and agent assistance.

[Cymbal Retail case study] Diagnostic Question #4

Cymbal Retail wants to use Generative AI to provide their customer service agents with real-time assistance. The system should listen to customer calls, transcribe them in real-time, and suggest relevant answers and solutions from a knowledge base.

Which combination of Google Cloud services would be best for this solution?

A. Speech-to-Text API, Vertex AI Language, and Cloud Storage.

B. Contact Center AI (CCAI) Platform, which integrates Speech-to-Text, Dialogflow, and Agent Assist.

C. Dialogflow CX, Speech-to-Text API, and BigQuery

D. Cloud Pub/Sub, Cloud Functions, and the Speech-to-Text API.

Google Cloud

B

The CCAI Platform is a purpose-built solution for this exact use case, providing a tightly integrated set of services for contact center automation and agent assistance.

[Cymbal Retail case study] Diagnostic Question #5

Cymbal Retail is building a product recommendation engine using Generative AI. The model should be able to recommend products based on a user's natural language query (e.g., "show me some stylish and comfortable shoes for a summer wedding").

Which approach would be most effective?

A. Use Vertex AI Search's semantic search capabilities to match user queries with relevant products.

B. Fine-tune a large language model (LLM) on Cymbal Retail's product catalog and customer reviews.

C. Use a traditional collaborative filtering model trained on user purchase data.

D. Build a custom model using TensorFlow and host it on Vertex AI Training.

Google Cloud

A

Vertex AI Search's semantic search capabilities are designed to understand the intent and context of natural language queries, making it a perfect fit for this use case.

[Cymbal Retail case study] Diagnostic Question #5

Cymbal Retail is building a product recommendation engine using Generative AI. The model should be able to recommend products based on a user's natural language query (e.g., "show me some stylish and comfortable shoes for a summer wedding").

Which approach would be most effective?

A. **Use Vertex AI Search's semantic search capabilities to match user queries with relevant products.**

B. Fine-tune a large language model (LLM) on Cymbal Retail's product catalog and customer reviews.

C. Use a traditional collaborative filtering model trained on user purchase data.

D. Build a custom model using TensorFlow and host it on Vertex AI Training.

Google Cloud

A

Vertex AI Search's semantic search capabilities are designed to understand the intent and context of natural language queries, making it a perfect fit for this use case.

Make sure to...

Enjoy the journey as much as the destination!

Now that you know about the overall setup of this course and how to use the workbook, let's get started by exploring section 1 of the exam guide.