398 million

398 million

Google Cloud mitigated the largest DDoS attack as of 2023, peaking at 398 million requests per second

What is Google Cloud?

Encryption at rest by default

115+ zones

Open Source

Security

Modern Infrastructure

Sustainable Tools and Infrastructure

35+ regions

Data and AI

Collaboration Cloud

What is Google Cloud?

Encryption at rest by default

Open Source

ern Infrastructure

35+ regions

AI

What is Google Cloud?

It's how you can...

Session 1
Host your first application

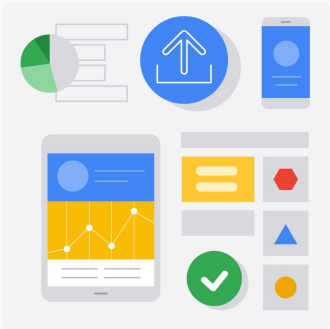
Session 2
Store and manage your data

Session 3
Analyze your data

Session 4
Get started with AI

Click to Deploy Solutions

- Move from business use case → deployment in **one-click**
- Covers a wide range of solutions from different pillars incl.
  - Application Modernisation
  - Data Analytics/Management,
  - AI
  - Security and Monitoring
- Reflects Google Cloud best practices and model architectures
- Will be used throughout this webinar to demo our solutions and for you to get started building!



goo.gle/loff-demo24-click-to-deploy

How to host your first application

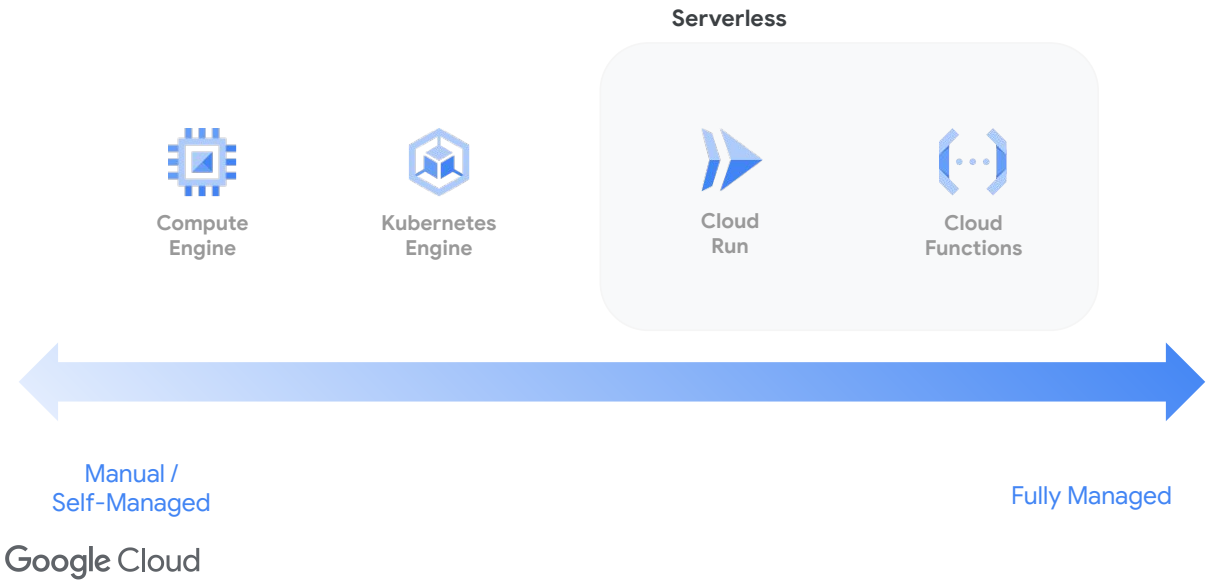
Google Cloud

What are you building?

Google Cloud

Step 1: Choose your platform

Compute on Google Cloud

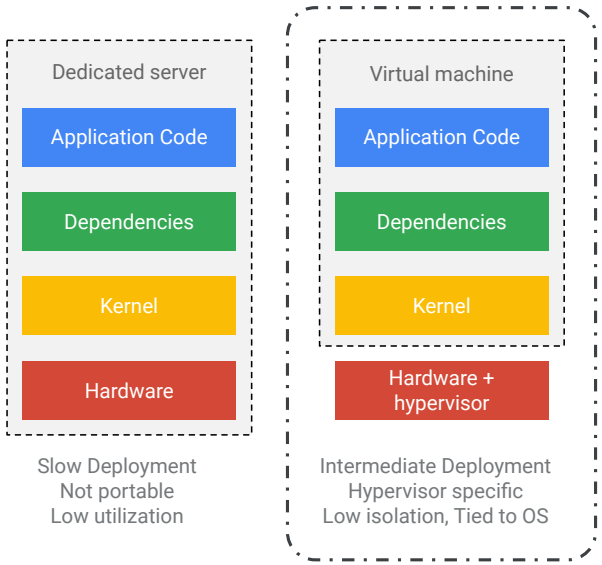


Compute Engine

What is a Virtual Machine?
A VM is a virtualized instance of a computer that can perform almost all of the same functions as a computer, including running applications and operating systems.

Virtual Machines on Google's infrastructure

- General purpose and optimized VMs
- Custom and predefined machine types
- Right sizing recommendations
- Live migration
- Committed and sustained use discounts
- Spot VMs, BYOL, and sole tenants

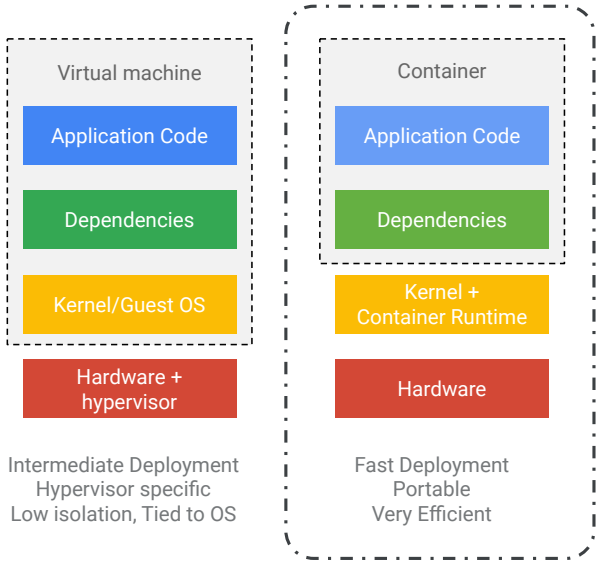


Kubernetes Engine

What is a Container?
Containers are lightweight packages of your application code together with dependencies such as specific versions of programming language runtimes and libraries required to run your software services.

Enterprise Container Management

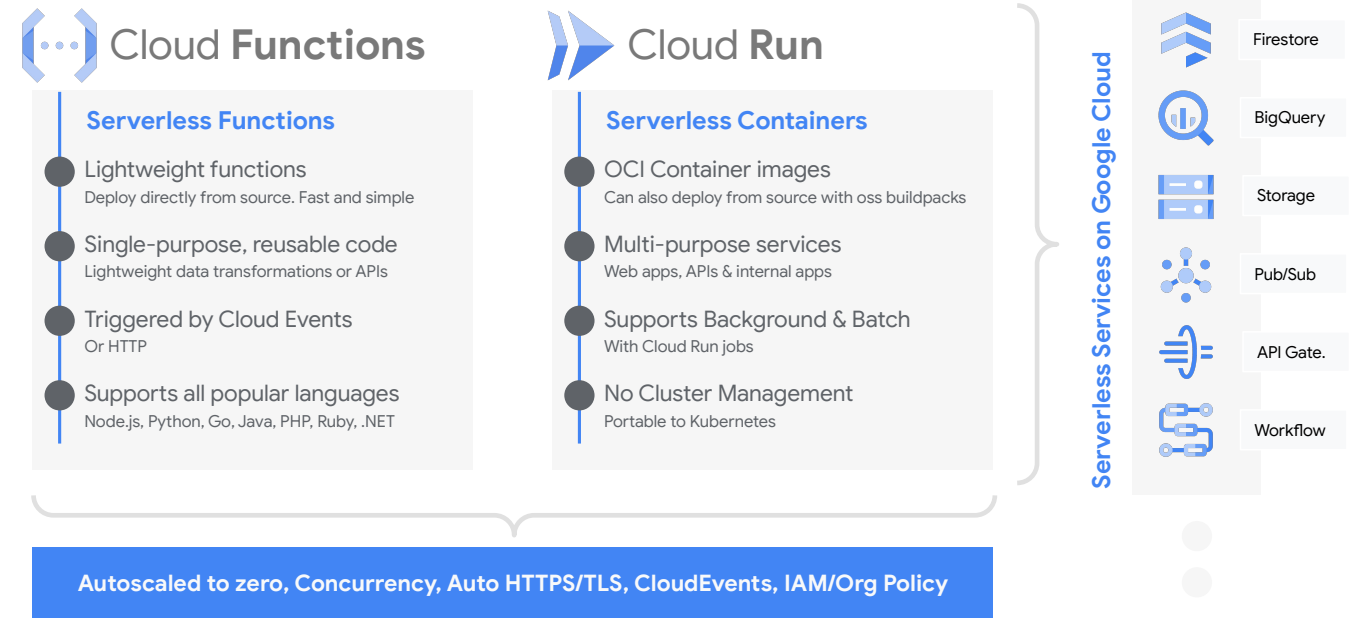
- **Scheduling**
  Decide where my containers should run
- **Lifecycle and health**
  Keep my containers running despite failures
- **Scaling**
  Make sets of containers bigger or smaller
- **Naming and discovery**
  Find where my containers are now
- **Load Balancing**
  Distribute traffic across a set of containers
- **Storage volumes**
  Provide data to containers
- **Logging and monitoring**
  Track what's happening with my containers
- **Debugging and introspection**
  Enter or attach to containers
- **Identity and authorization**
  Control who can do things to my containers



Intermediate Deployment
Hypervisor specific
Low isolation, Tied to OS

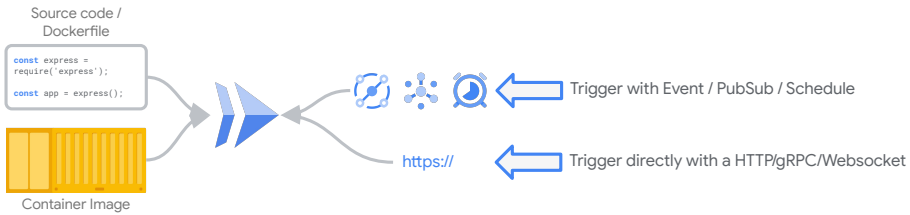
Fast Deployment
Portable
Very Efficient

Serverless Portfolio



Google Cloud

Cloud Run



Web Services, APIs, Microservices

- **Optimized for developer productivity**
  Fast time to market - minimal configuration and no knowledge of Kubernetes necessary
- **Automatically managed & scaled**
  Hyper-elastic scaling, including scale to zero
- **No Cluster Management**
  No infrastructure to manage, portable to Kubernetes
- **Deploy source, Dockerfile or image**
  Uses open source buildbacks and Cloud Build, integrates with Artifact Registry or Docker Hub

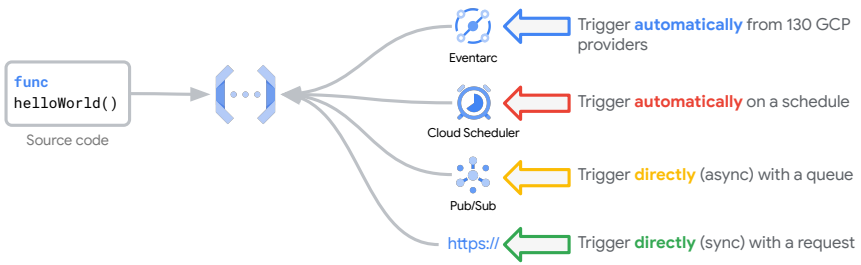
Pricing

Perpetual Free Tier
2M requests, 180K vCPU-seconds
Per month

Pay per second of execution time
Price varies based on memory & cpu allocation. HTTP included with request fee. Request fees waived for Always-on CPU

Google Cloud

Cloud Functions



Event-based applications, data transformations, ops automation

- **Deploy source code**
  Online editor, from bucket, using Github Actions
- **Bind to trigger**
  Automatically on event, or directly sync & async
- **Auto-scales down to zero**
  No pre-provisioning, only pay for what you use
- **Supports all popular languages**
  Node.js, Python, Go, Java, PHP, Ruby, .NET

Pricing

Perpetual Free Tier
2M requests, 180K vCPU-seconds
Per month

Pay per second of execution time
Price varies based on function size
HTTP included with request fee

Google Cloud

Fill in the blank

Compute Engine

_____ _____ on Google's infrastructure

Web apps, databases, machine learning, etc

Cloud Run

_____ containers managed by Google

Web apps, APIs, microservices

Kubernetes Engine

_____ management service from Google

Web apps, databases, machine learning, etc

Cloud Functions

Serverless _____ managed by Google

Event-based applications, data transformations, ops automation

Fill in the blank

Compute Engine

Virtual machines on Google's infrastructure

Web apps, databases, machine learning, etc

Cloud Run

Serverless containers managed by Google

Web apps, APIs, microservices

Kubernetes Engine

Container management service from Google

Web apps, databases, machine learning, etc

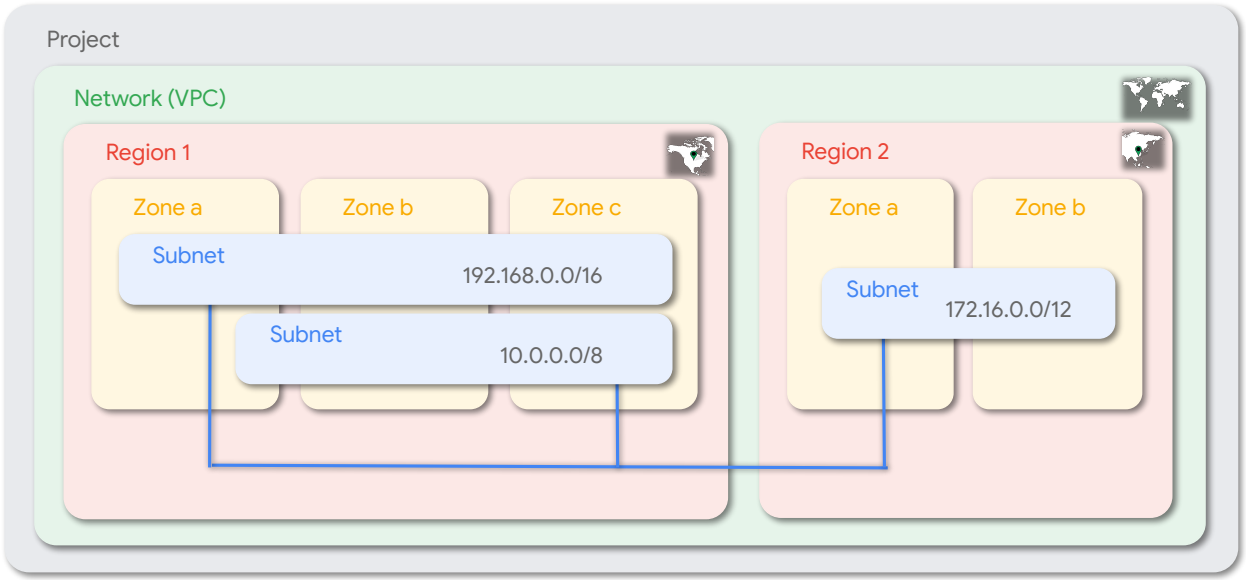
Cloud Functions

Serverless functions managed by Google

Event-based applications, data transformations, ops automation

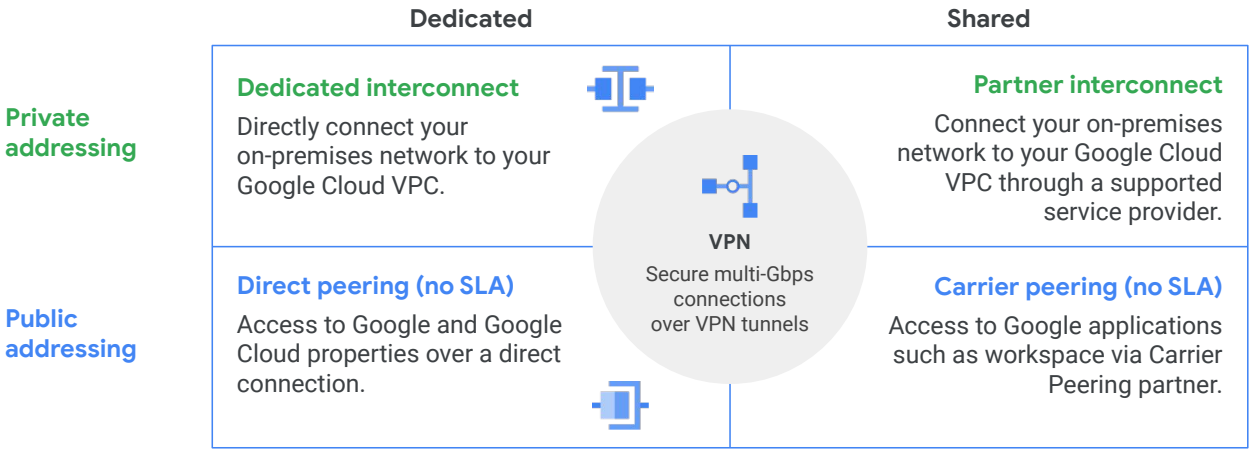
Step 2: Connect your platform

Google VPC



Google Cloud

Hybrid and multi-cloud connections



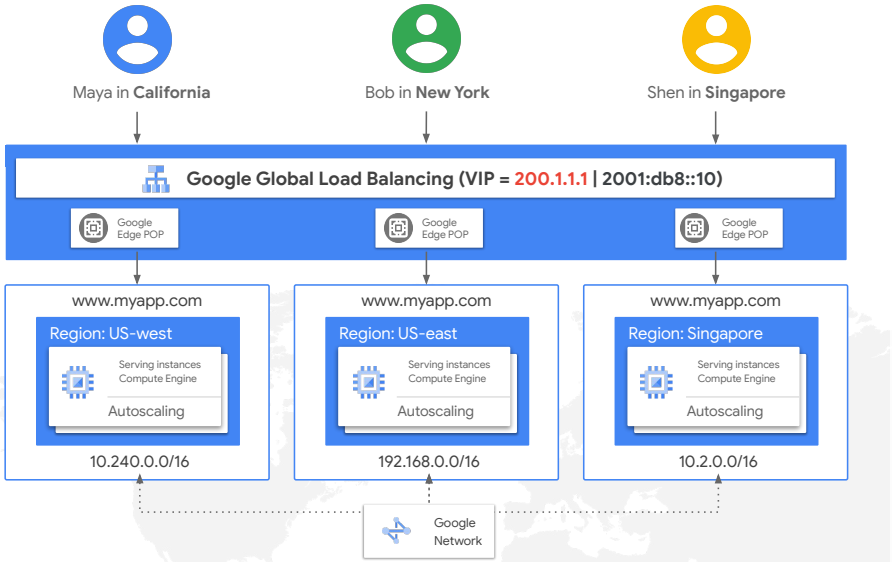
Google Cloud

Load balancer types

	Type	Geographical scope	Network tiers	Proxy/pass-through
Internal	TCP/UDP	Regional	Premium	Pass-through
	HTTP(s)			Proxy
External	TCP/UDP	Regional	Standard / Premium	Pass-through
	HTTP(s) TCP Proxy SSL Proxy	Regional / Global depending on network tier	Standard / Premium	Proxy

Google Cloud

Serving Application Traffic: External global HTTP(s) load balancing

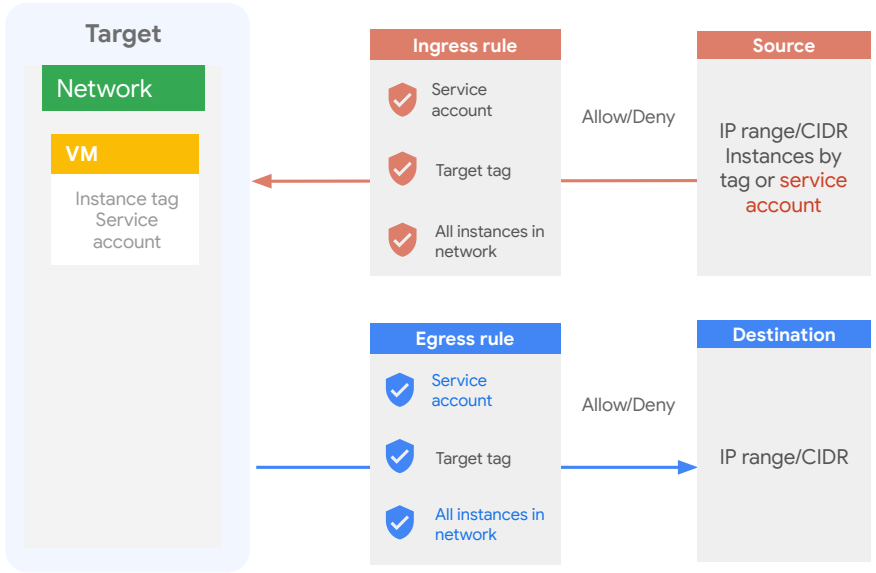


Google Cloud

- Global anycast IP address
- Managed service on Google Front Ends (GFE)
- Multi-regional load balancing
- Balancing mode: RATE, UTILIZATION
- Session affinity: Client IP, generated cookie, header field, & HTTP Cookie
- WebSocket support
- Terminates HTTPs traffic as close a possible to users.
- Backend traffic encryption supported.
- DDOS protection (L4)
- Integration with Cloud CDN and Cloud Armor (WAF)

Step 3: Secure and monitor

VPC firewall



Google Cloud

VPC firewall

- **Stateful** with connection tracking
- **Distributed**: enforced on underlying host

Control paths

- VM <-> VM
- VM <-> Internet
- VM <-> On-prem

Implied rules

- Ingress deny
- Egress allow

Attaching firewall rules to VMs

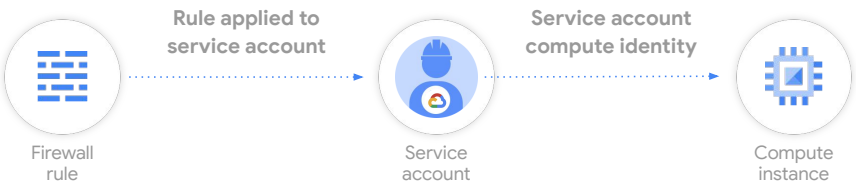
Network Tags

- Multiple tags applied to one VM (64 max)
- Firewall rule may target multiple tags
- May update tags to live VM

Best practice

Service accounts

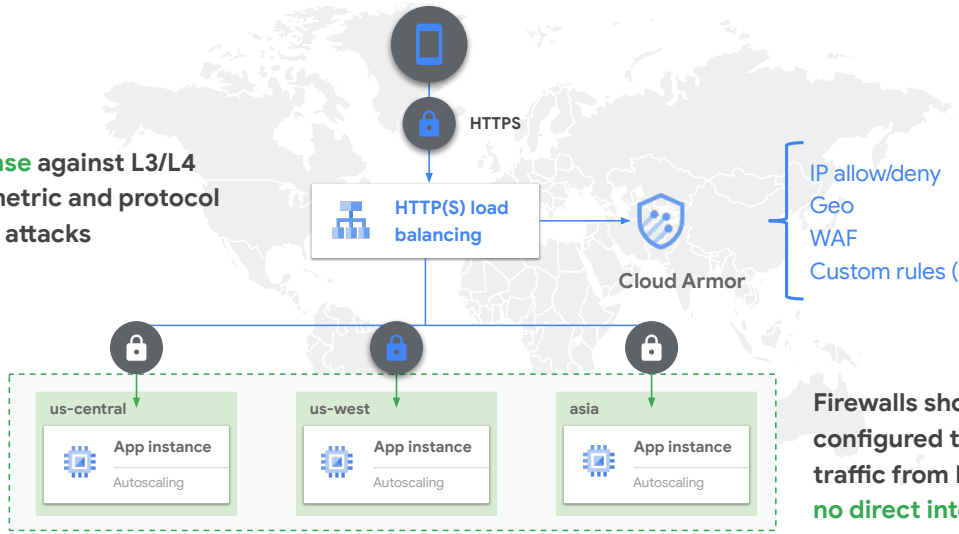
- May restrict who uses
- Must shut down VM to change service account



Google Cloud

Cloud Armor: DDoS protection and WAF

Defense against L3/L4 volumetric and protocol DDoS attacks



Firewalls should be configured to only allow traffic from HTTP(S) LB, no direct internet traffic

Google Cloud

Cloud Operations Suite

Logging and Error Reporting

Cloud Logging

- Fully managed, real-time log management with storage, search, analysis and alerting at exabyte scale.

Error reporting

- A single place to monitor error conditions from all apps and services

Monitoring

Cloud Monitoring

- Gain visibility into the performance, availability, and health of your applications and infrastructure.

Application Performance Management

Cloud Trace

- Find performance bottlenecks in production.

Cloud Profiler

- Identify patterns of CPU, time, and memory consumption in production.

Google Cloud

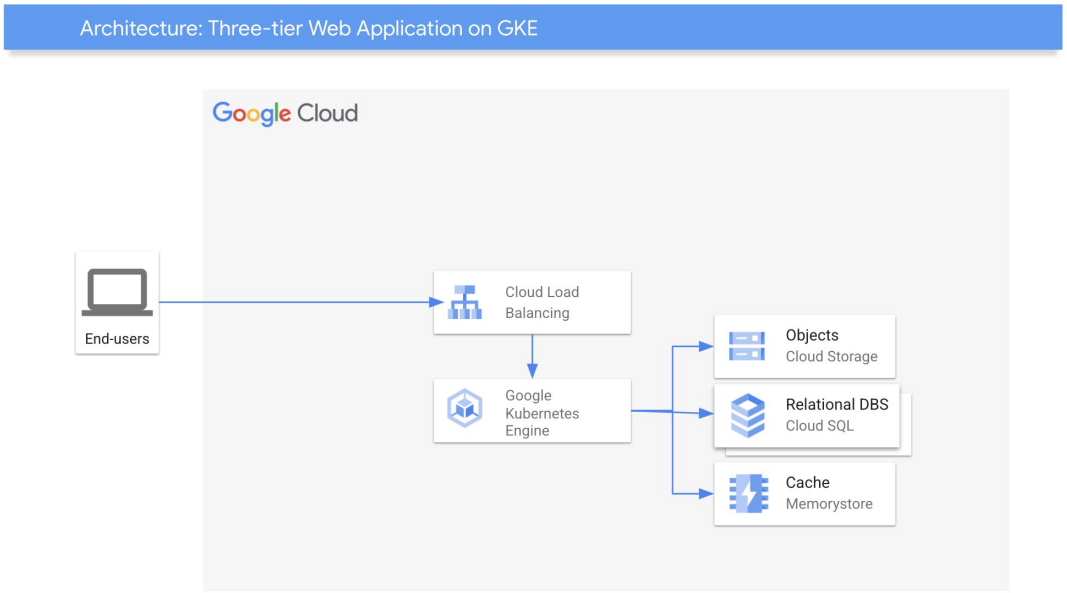
Putting it all together

Three-tier web application with Google Kubernetes Engine (GKE)

Example of a three-tier web application with Google Kubernetes Engine (GKE).

The application's services run on a GKE cluster, enabling autoscaling if necessary.

Cloud SQL is used as relational database, Cloud Storage for objects. Memorystore is used for caching to reduce access to the database on frequent queries.



Try it yourself

Demo :
goo.gle/loff-demo24-TTGKE

More solutions:
goo.gle/loff-demo24-click-to-deploy

What about data?

Google Cloud

How to store and manage your data

Google Cloud

What types of data are you working with?

Google Cloud

Managed database options

Cloud SQL	AlloyDB	Spanner	Bigtable	Firestore
Enterprise-ready fully managed relational database service for PostgreSQL, MySQL, SQL Server	PostgreSQL-compatible database ready for enterprise level workloads	Global scale and 99.999% availability with PostgreSQL interface	Key-value database with flexible schema, single-digit millisecond low latency, and high throughput	Serverless document database with a rich development ecosystem and backend as a service
Modernize with fully managed databases	**High performance** with open source compatibility and the best of Google	**Scalability and availability** for the demanding workloads	**Fast reads and writes** for high throughput workloads	**Native integration** with Google Cloud and Firebase

Cymbal Bank

Google Cloud

VP of IT

Adityah

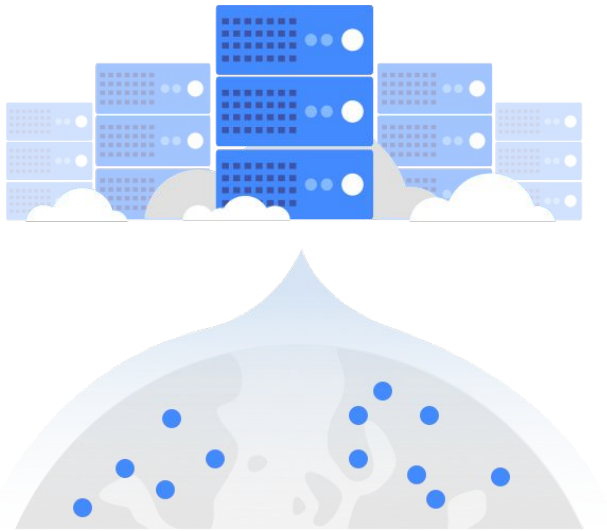
Barriers to overcome

- Operational complexity
- Lack of monitoring & troubleshooting
- Security & cost risks
- Complicated migrations

Cloud SQL

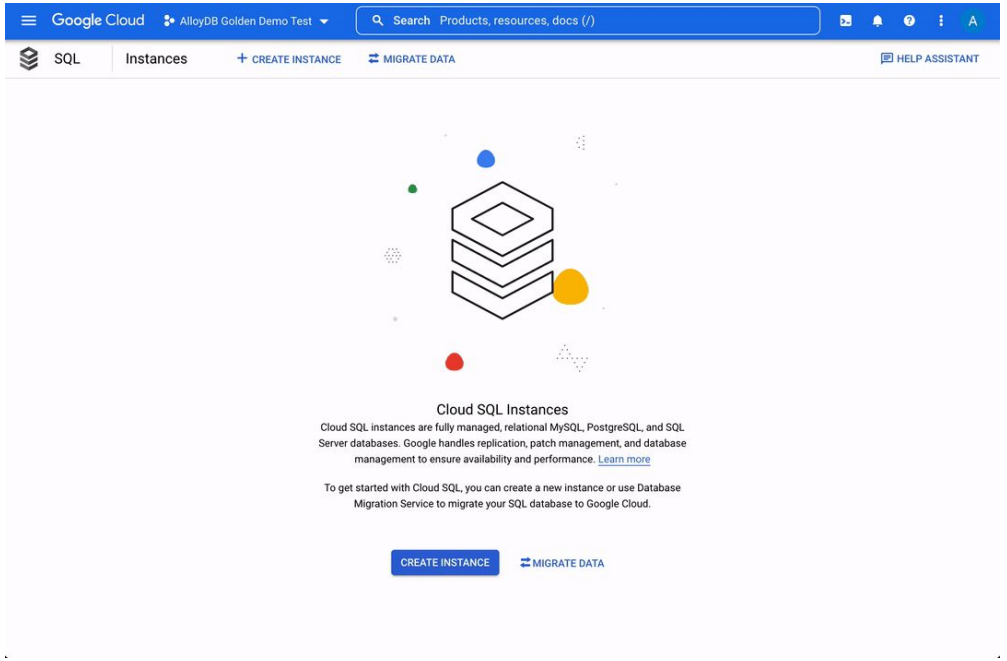
Fully managed relational database service

Support for MySQL PostgreSQL Microsoft SQL Server



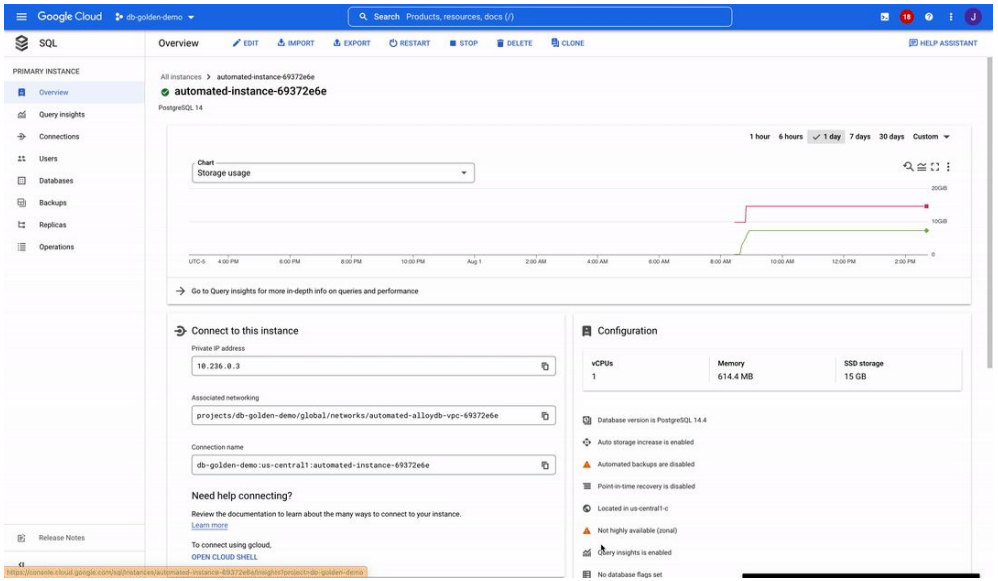
Automated and fully managed

Provisioning



Automated and fully managed

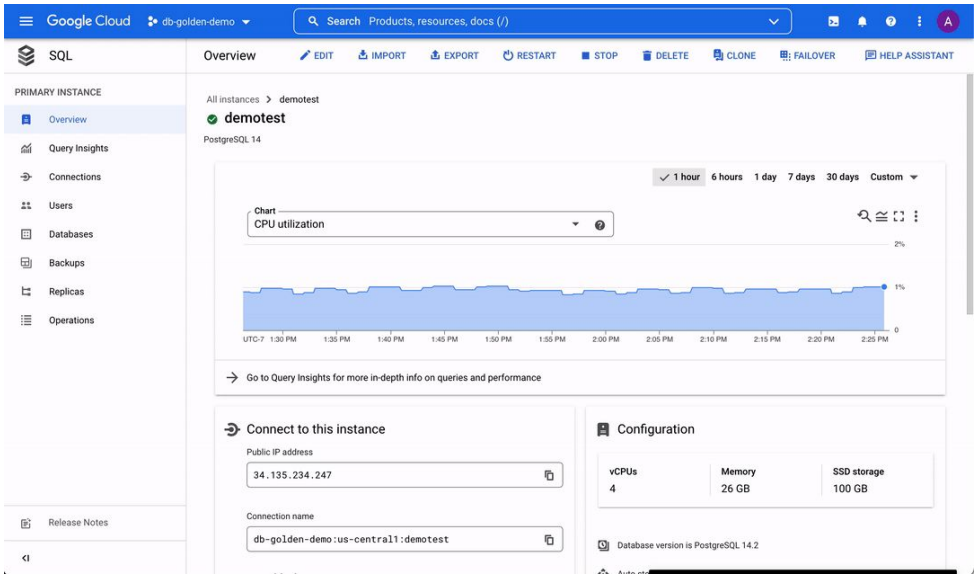
Auto storage increase



Google Cloud

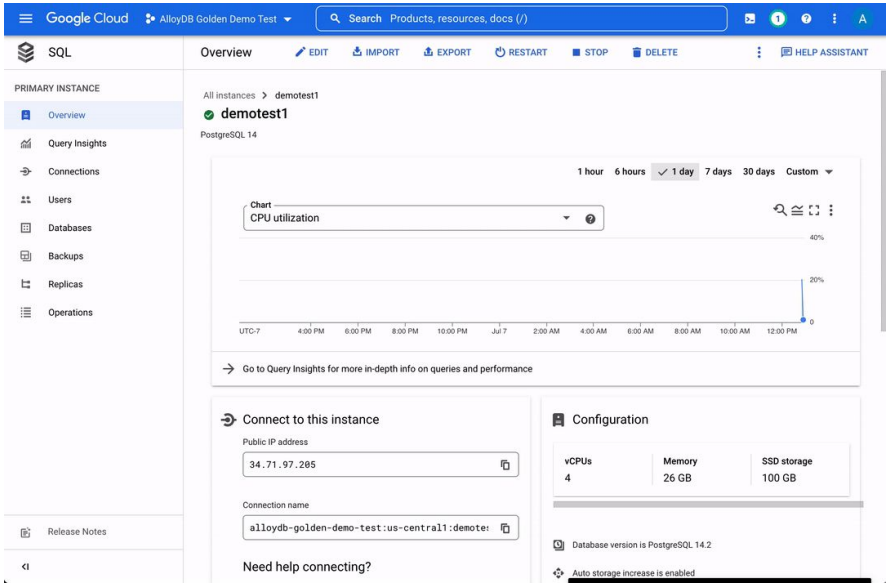
Automated and fully managed

Connectivity



Automated and fully managed

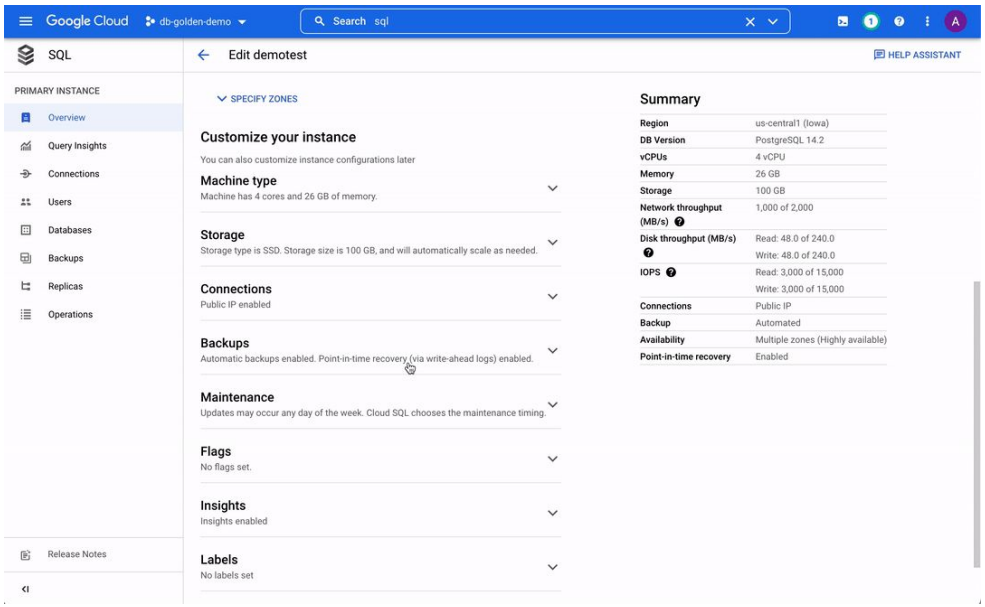
Backup and recovery



Google Cloud

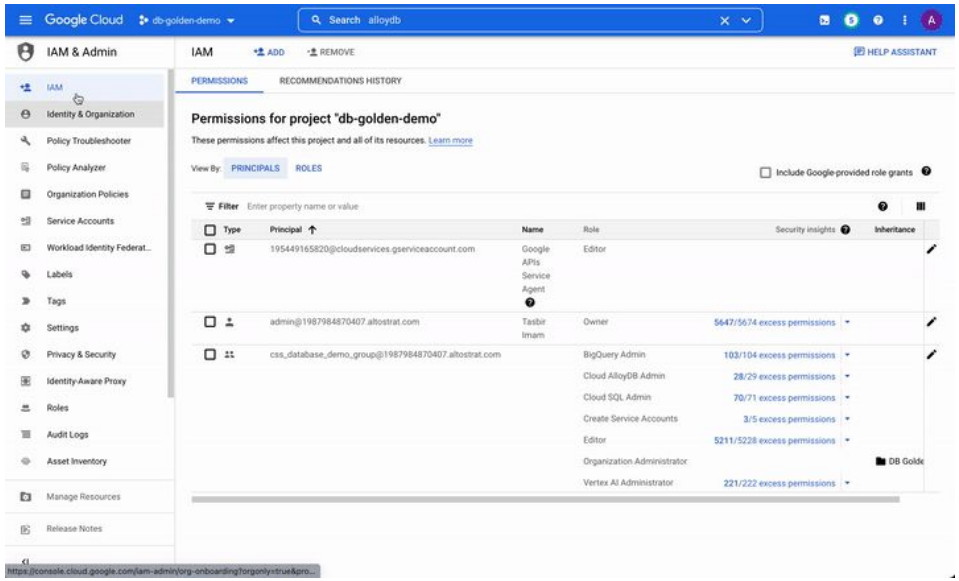
Automated and fully managed

Maintenance



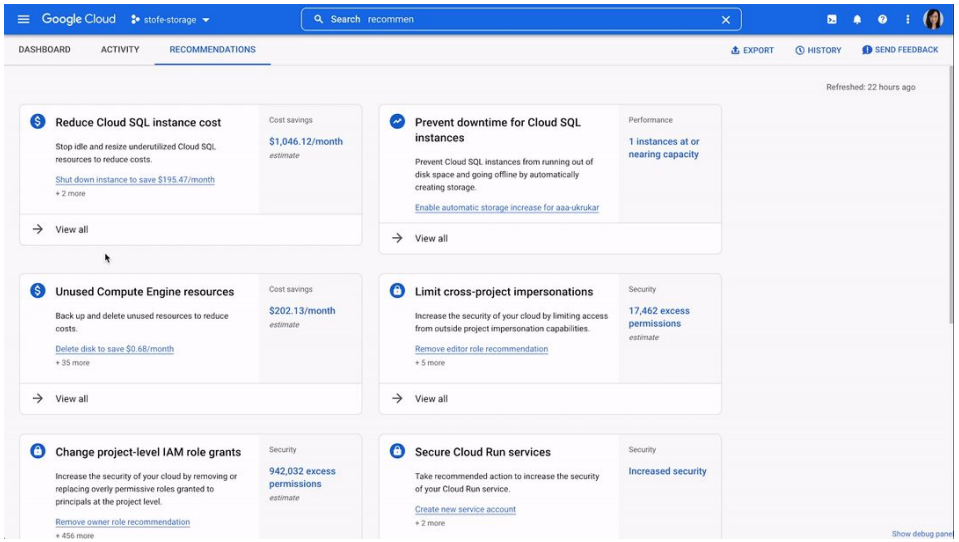
Risk Reduction

Security



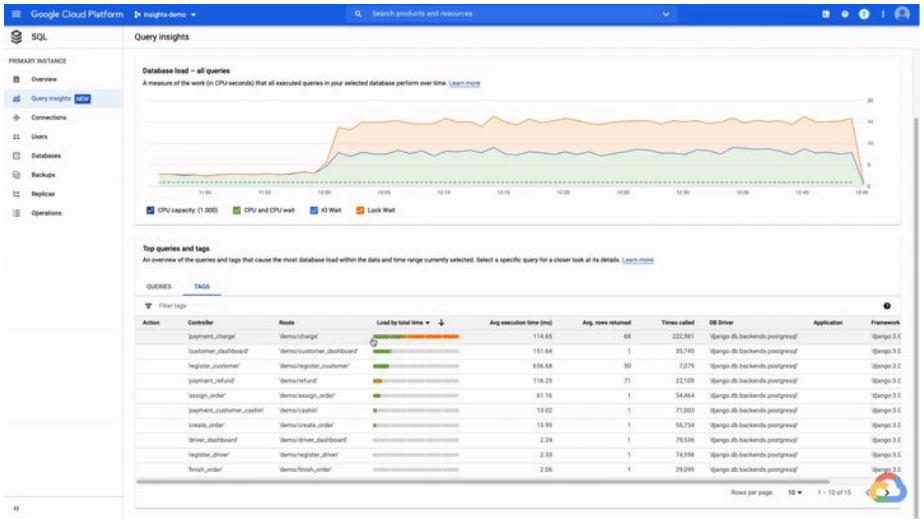
Risk Reduction

Cost Recommender



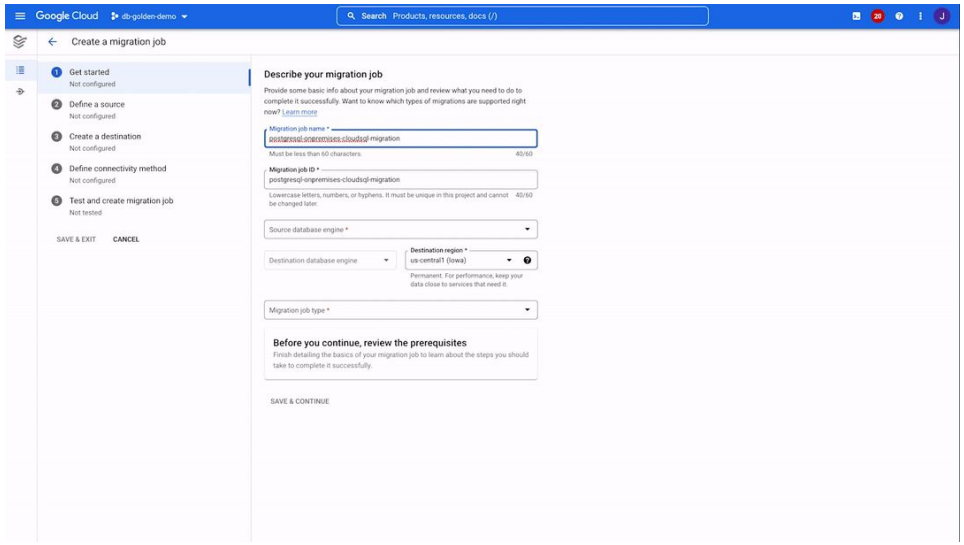
Productivity and observability

Cloud SQL Query Insights tool provides monitoring and diagnostics that helps to detect and fix query performance problems.



Easy Migration

Migration to Cloud SQL using Database Migration Service is serverless, and easy to use.

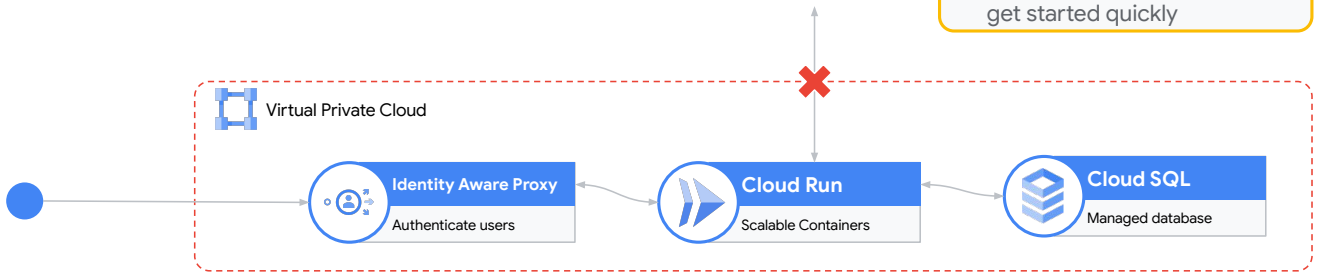


Design patterns

Internal corporate apps

Advantages

Internal apps which don't receive high volumes of traffic can benefit from Cloud Run and Cloud SQL to get started quickly



Your private VPC may contain internal resources and users with a strict security boundary enforced at the network level

Authenticate your users with **Identity Aware Proxy**

Cloud Run will only accept requests from within the VPC. Cloud Run is serverless and can scale up and down according to workload traffic.

Cloud SQL provides fully managed open source and commercial database options

VP of IT

Adityah

Cloud SQL

- Fully managed operations
- Risk reduction and governance
- Better observability
- Easy migration

Cymbal Bank's **new initiative**

Offering customers a unified experience for both banking and investment account management

Head of Digital Banking

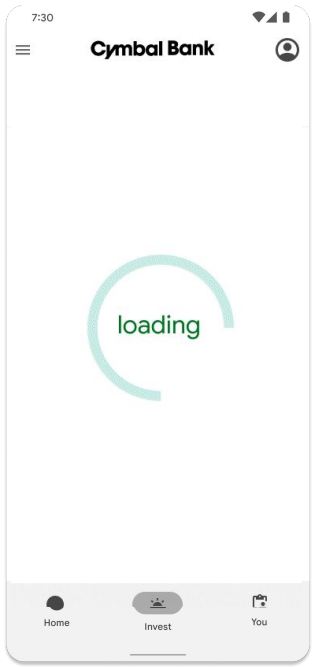
Elena

- Increasing trade volumes and volatility
- Need real-time insights for my customers

Cymbal Bank

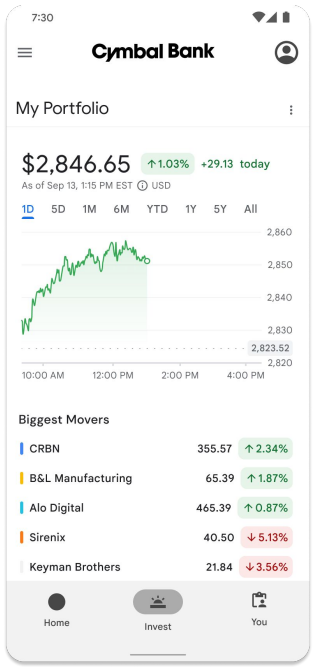
Real-time Operational Insights

AlloyDB vs. Conventional PostgreSQL:
A side-by-side comparison



Status quo

VS



w/ AlloyDB

Google Cloud

AlloyDB

2x faster

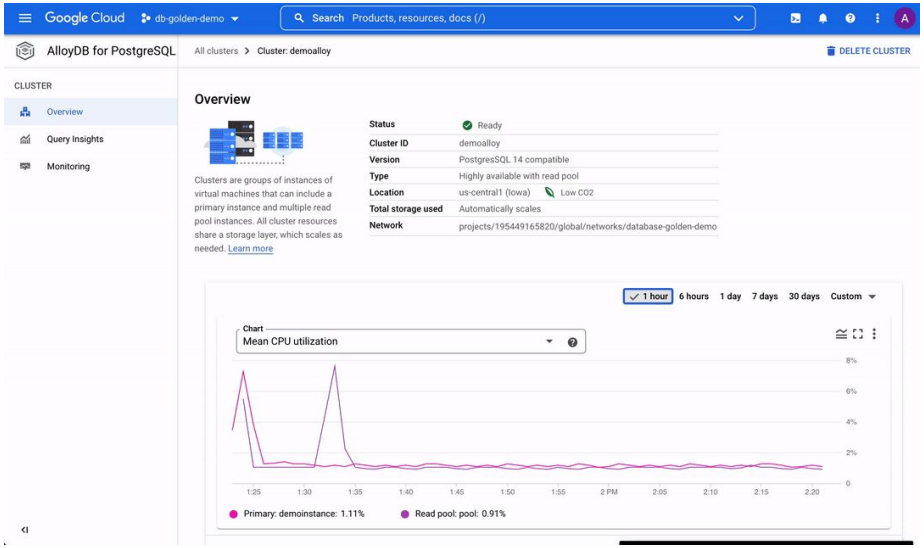
than Amazon's comparable PostgreSQL-compatible service for transactional workloads

100x faster

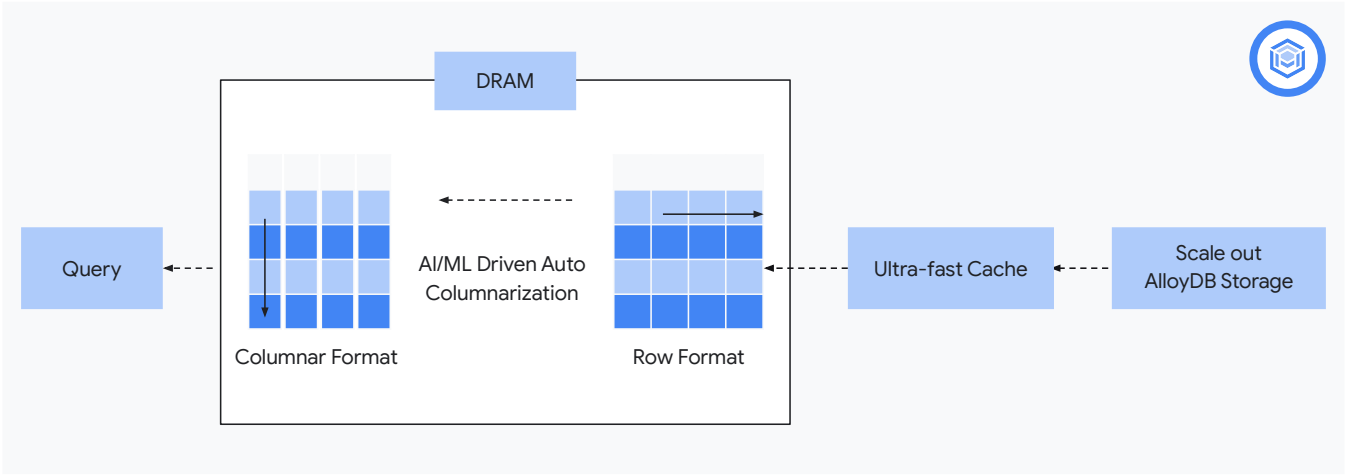
for analytical queries than standard PostgreSQL

Real Time Operational Analytics

Hybrid transactional and analytical processing (HTAP)



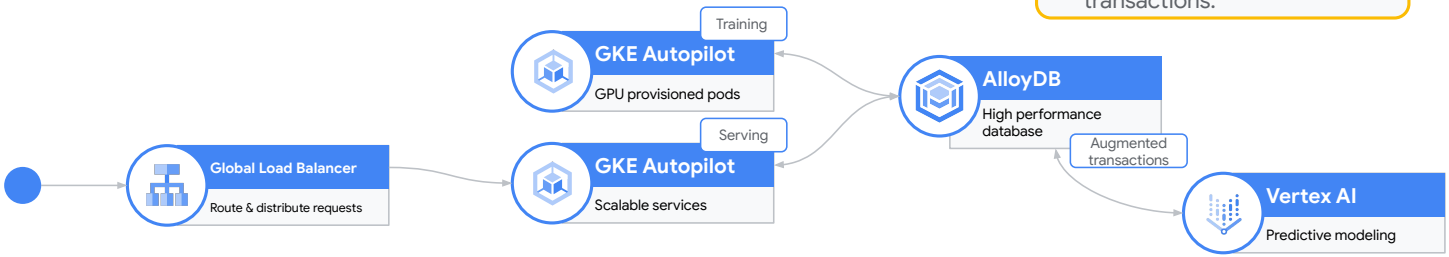
The columnar cache that's powering the dashboard is itself powered by ML based on the unique characteristics of the workload



Google Cloud

Design patterns

AI/ML augmented transactions



Advantages

Use GKE Autopilot to easily provision infrastructure like GPUs. AlloyDB serves ML models through augmented transactions.

Serving

GKE Autopilot provides highly scalable, low-overhead managed Kubernetes for serving ML workloads

Easy GPU access

GKE Autopilot provides easy "on demand" provisioning of hardware accelerators like GPUs for use cases like AI/ML and protein folding

Data processing

AlloyDB is a PostgreSQL-compatible database for ML serving and augmented transactions

Google Cloud

Head of Digital Banking

Elena

AlloyDB

Trade volume scaling without issues

New data value created for
Cymbal Bank and their customers

Predictable and transparent pricing

Head of Digital Banking: Initiative #2

Elena

A Specific Technical Challenge:

Digital trading app uses HBase for delivering real time recommendations to the user and for detecting fraud

HBase is not scaling to the needs of the app - outages occur during heavy trading periods

HBase is costly and difficult to maintain

Why modernize with Bigtable?

- Bigtable is a managed NoSQL service that supports the open HBase API - **no code lock in**
- We provide the tools to help automate assessment and subsequent migration **easier and faster**
- We have migrated **hundreds of customers** across every industry
- **Google uses Bigtable** extensively in its most popular apps

Cymbal Bank

Bigtable Common Use Cases

Personalization and Recommendations with real-time clicks and transactions

Fraud Detection with real-time events

Predictive Maintenance with IoT / Machine Data

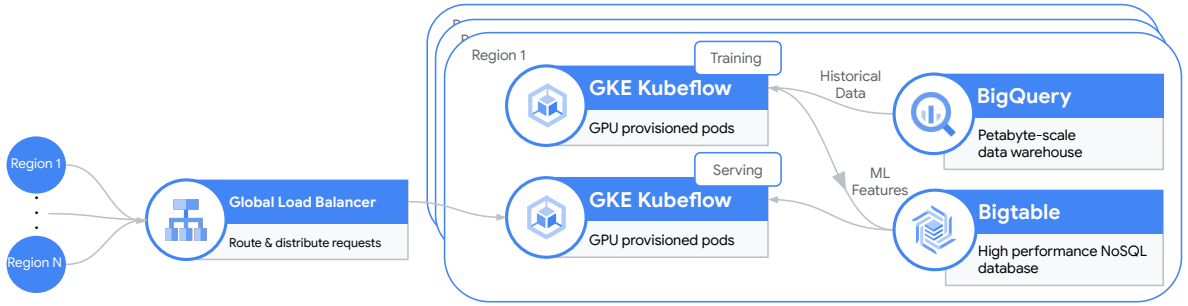
Metadata Management for customer and product metadata

Design patterns

AI/ML training and low-latency serving

Advantages

Use GKE and Bigtable for workloads that require low-latency and high throughput



ML workflow	Historical analysis	Online inference & model serving	Global deployments
Kubeflow on GKE simplifies the build-train-deploy lifecycle.	**BigQuery** provides the historical data for analysis, feature engineering, offline model training and evaluation.	**Bigtable** provides high read & write throughput with low latency. It can adapt to fluctuating demand with built-in autoscaling. Couple with **MLRun, Tensorflow Serving or KServe on GKE**, to serve real-time predictions at scale for ML framework of your choice.	**Bigtable** supports read/write clusters in **up to 8 regions of your choice** for fast regional access, higher resiliency and data residency compliance.

Head of Digital Banking

Elena

On premise HBase replaced with fully managed Bigtable

Automatic scaling up and down to meet the needs of the app - no more capacity related outages

Cymbal Superstore

Retailer with a large,
rapidly growing digital business

Director of Ecommerce

Elma

Online store application has outages during heavy use periods - no managed database product scales well enough

Application must run in multiple regions, but it must look like it is one database - we must have a consistent view of all data

Even an outage for a few minutes (planned or unplanned) is too expensive - want 99.999% SLA

Cloud Spanner

Not just another _____ database.

Performance

and

Scale

and

SQL

and

Fully Managed Service

Cymbal Superstore

What is Unique About Spanner?

Battle tested by Google

Relational semantics

Schemas, ACID transactions, SQL

+

Horizontal scale

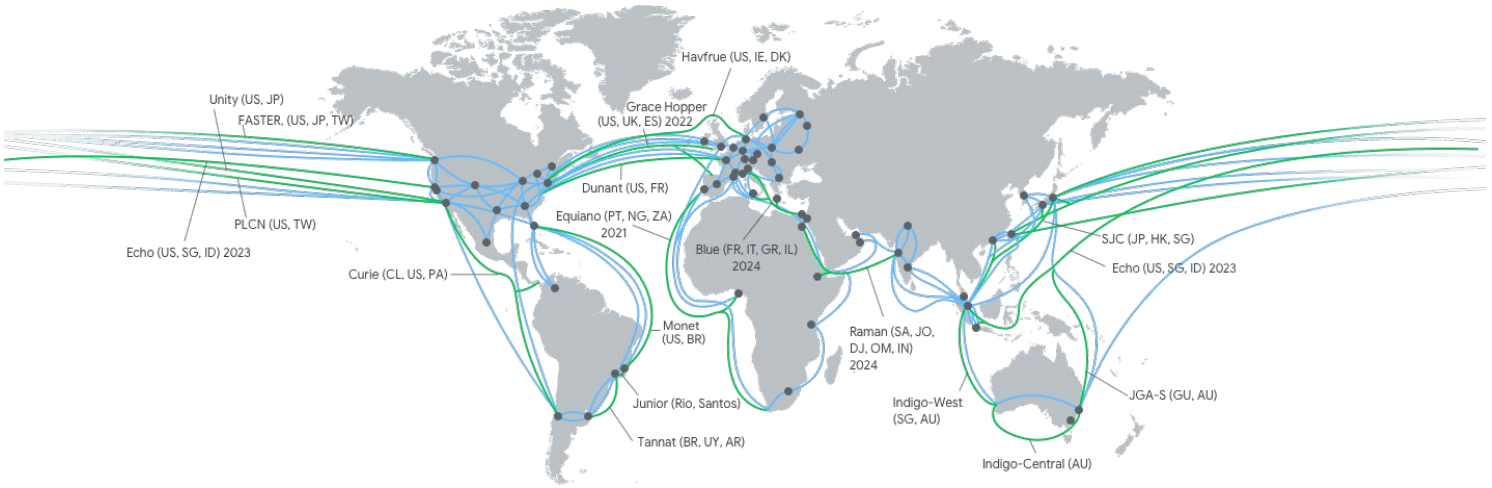
99.999% SLA, fully managed, and scalable

Spanner is a fully managed database service that processes over 2 billion requests per second at peak

Cymbal Superstore

Google's Unique Worldwide Network

Powered by Google



Google Cloud

Cymbal Superstore

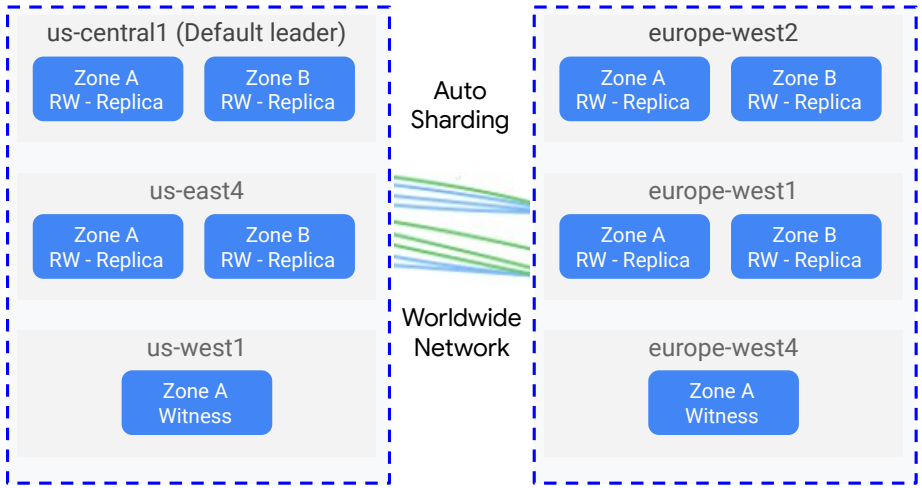
Availability at any scale

Industry leading 99.999% SLA
RPO = 0, RTO = 0

Multiple read-write regions with a consistent, single view of all data

No maintenance window

Online schema changes



Multi region example

Google Cloud

Cymbal Superstore

Spanner Continued Innovation

Democratized Access

Reducing the barriers for teams to access Spanner's unique benefits:

- Lower entry price: Granular instances
- Committed use discounts: Up to 40%
- Familiarity, portability: PostgreSQL interface
- Free Trial Instance

Enterprise Readiness

Powering mission-critical applications in regulated industries:

- Point-in-time recovery
- CMEK organization policy
- Fine-grained access control
  PREVIEW

Developer Experience

Simplifying integration into applications and developer workflows:

- Query Insights
- BigQuery federation
- Cost-based query optimizer
- Self-service optimizer stats updates (`ANALYZE`)
- Open Telemetry, Grafana support

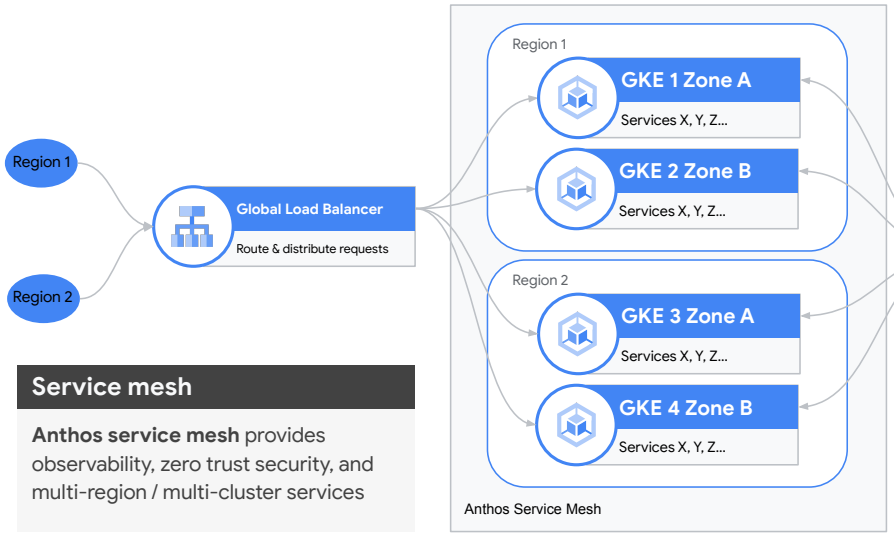
Performance and Flexibility

Realizing the benefits of the cloud for operational workloads:

- Doubled storage: 4TB/node
- Mutation limit: 40k
- Leader placement
- New multi-regions

Design patterns

Business continuity for enterprise apps



Service mesh

Anthos service mesh provides observability, zero trust security, and multi-region / multi-cluster services

Advantages

Use Spanner for workloads that need to automatically failover with high reliability and high consistency

Kubernetes at scale

GKE Fleets eases configuring and running regional services for high availability and business continuity

Database

Use **Cloud Spanner** to provide a globally-consistent SQL database with 99.999% availability

Google Cloud

Firestore's Mission

Unlock application innovation with **simplicity**, **speed**, and **confidence**.

What makes Firestore special?

Serverless document database

Built for developers: scalable, schemaless (e.g. JSON), ACID transactions, pay only for what you use

Backend as a service

Connect directly, no middle tier needed, development to production in days

Built-in real-time sync & offline mode

Easily build rich apps with live sync, that also fallback to offline caching when a client device loses network connectivity

Firestore has more than **250,000** monthly active developers

Differentiated infrastructure

01 **Serverless:** Google battle-tested, automic, elastic scaling, that helps minimize capacity planning. You pay for what you use.

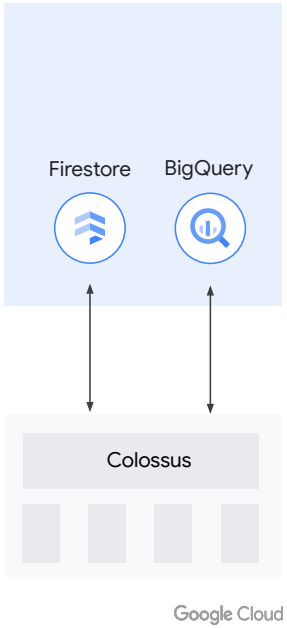
02 **Serializable transactions:** serializable ACID transactions with up to 500 documents.

03 **High availability:** up to 99.999% SLA availability (multi-region), 0 RPO/RTO for zonal and regional failures, and no maintenance window downtimes.

04 **External consistency:** external consistency based on TrueTime.

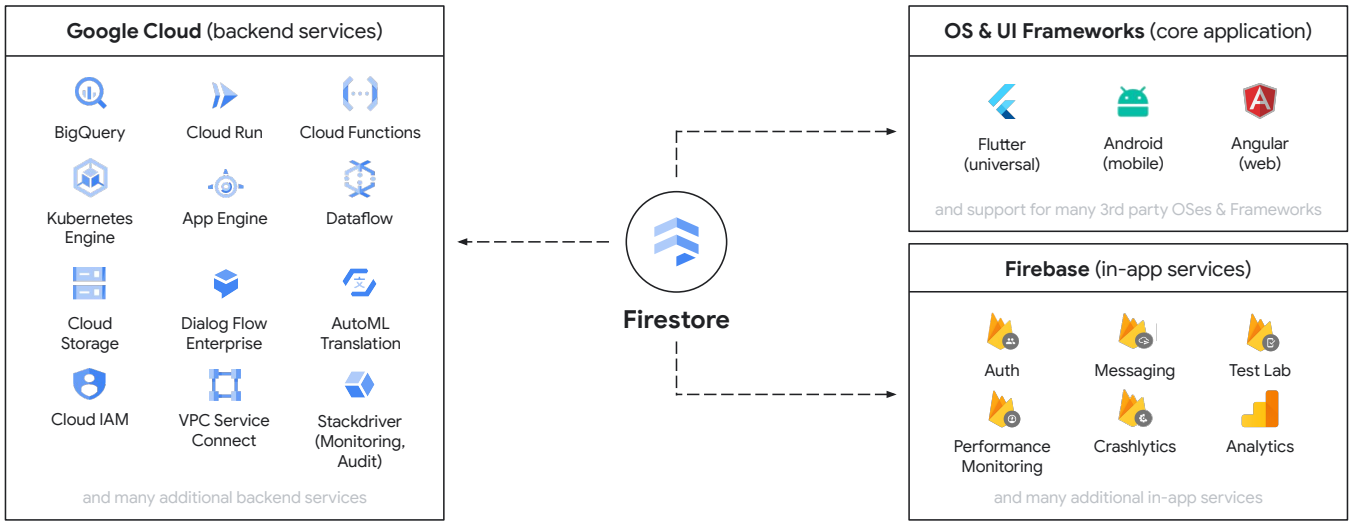
05 **Multi-version concurrency control:** allows customers to perform queries both on most recent and past data.

06 **Shared Colossus storage:** performant data access for load, storage and retrieval for a variety of use cases - OLTP, analytics, AI/ML



Google Cloud

Integrated full-stack application development ecosystem from Google



Director of Ecommerce

Elma

Massively scaling application even during peak times of heavy load. No outages due to capacity

Data resides in multiple regions, but consistent queries and reports look like it is one database

No data loss - no downtime for maintenance - instant recovery from failures

Managed database options

Enterprise-ready fully managed relational database service for PostgreSQL, MySQL, SQL Server	PostgreSQL-compatible database ready for enterprise level workloads	Global scale and 99.999% availability with PostgreSQL interface	Key-value database with flexible schema, single-digit millisecond low latency, and high throughput	Serverless document database with a rich development ecosystem and backend as a service
Modernize with fully managed databases	**High performance** with open source compatibility and the best of Google	**Scalability and availability** for the demanding workloads	**Fast reads and writes** for high throughput workloads	**Native integration** with Google Cloud and Firebase

Managed database options

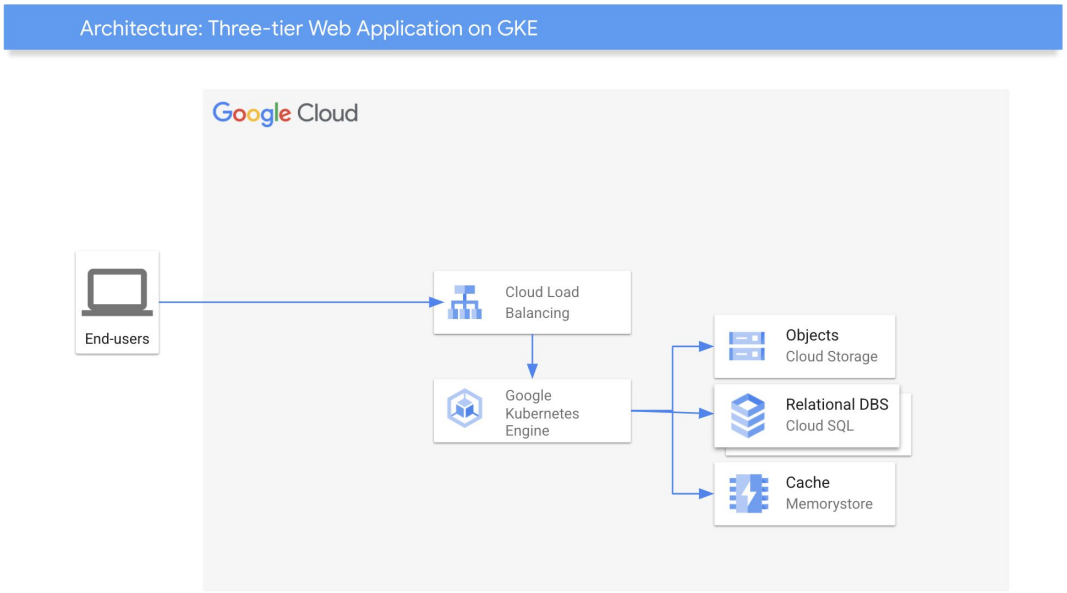
Cloud SQL	AlloyDB	Spanner	Bigtable	Firestore
Enterprise-ready fully managed relational database service for PostgreSQL, MySQL, SQL Server	PostgreSQL-compatible database ready for enterprise level workloads	Global scale and 99.999% availability with PostgreSQL interface	Key-value database with flexible schema, single-digit millisecond low latency, and high throughput	Serverless document database with a rich development ecosystem and backend as a service
Modernize with fully managed databases	**High performance** with open source compatibility and the best of Google	**Scalability and availability** for the demanding workloads	**Fast reads and writes** for high throughput workloads	**Native integration** with Google Cloud and Firebase

Three-tier web application with Google Kubernetes Engine (GKE)

Example of a three-tier web application with Google Kubernetes Engine (GKE).

The application's services run on a GKE cluster, enabling autoscaling if necessary.

Cloud SQL is used as relational database, Cloud Storage for objects. Memorystore is used for caching to reduce access to the database on frequent queries.



Try it yourself

Demo :
goo.gle/loff-demo24-TTGKE

More solutions:
goo.gle/loff-demo24-click-to-deploy

How to analyze your data

Google Cloud

What is a data warehouse?

Google Cloud

What is a data warehouse?

A data warehouse is an enterprise system used for the **analysis and reporting of structured and semi-structured data from multiple sources**, such as point-of-sale transactions, marketing automation, customer relationship management, and more.

Google Cloud

Google BigQuery

Google Cloud Platform's **enterprise data warehouse** for analytics

Gigabyte- to **petabyte-scale** storage and SQL queries

Encrypted, durable, And highly available

Fully managed and **serverless** for maximum agility and scale

Unique

Real-time insights from streaming data

Unique

Built-in **ML** for out-of-the-box predictive insights

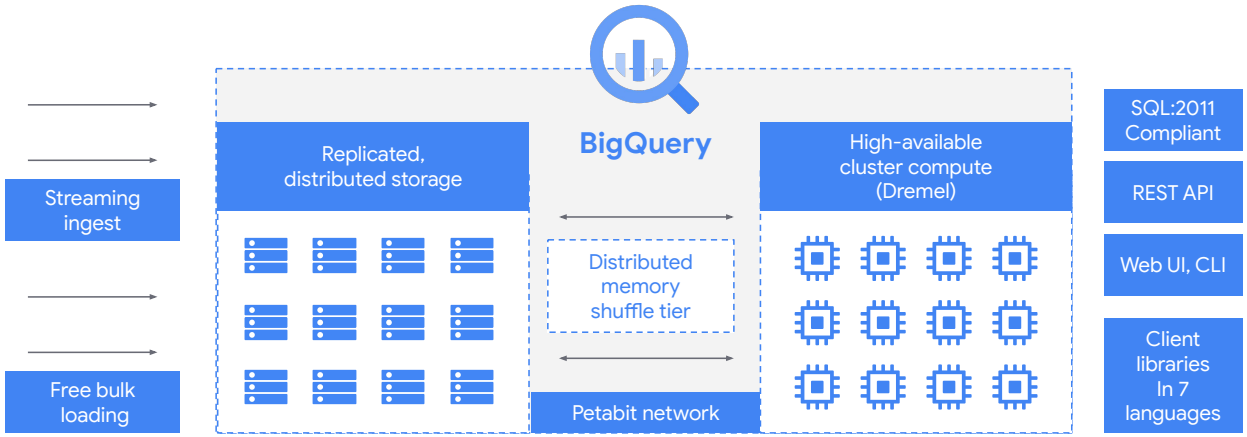
Unique

High-speed, in-memory **BI Engine** for faster reporting and analysis

Unique

Google Cloud

BigQuery Architecture



Google Cloud

Completely serverless

Why BigQuery?

- Simplifies capacity management
- Dynamically adjusts to demand

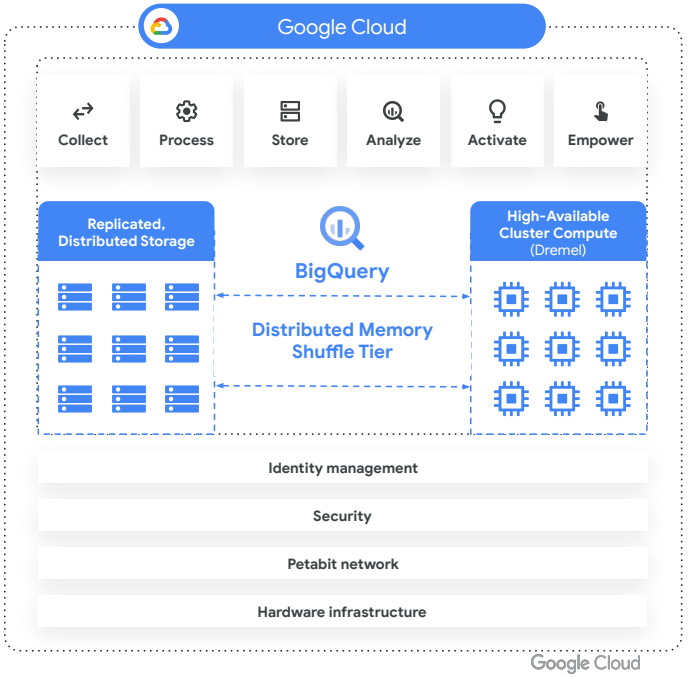
VS

- Plan, manage, pay VMs
- Limit use data due to capacity restrictions

Elastic Computing

Distributed storage and compute with high bandwidth including distributed petabyte scale in-memory storage for temp data and state:

- Auto-start and auto-pause
- 0-Second warm up to get maximum performance
- Accelerate queries in flight



Stream Analytics

Why BigQuery?

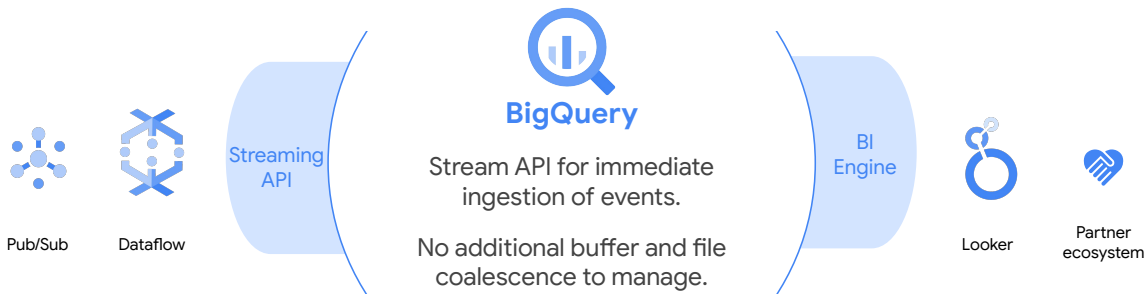
- Leverage event-driven analysis with built-in streaming capabilities

VS

- Leverage historical analysis and batch processing

Analyze business events in real time

Move your business to event-driven action for logstream, clickstream, and sensor data to enable use cases like anomaly detection and continuous intelligence.



Built-in AI/ML | BQML

Why BigQuery?

- Provide ML access to more users through a simple SQL interface

VS

- Require every ML use case to go through more specialized systems that require advanced skill sets

Machine Learning for all Built-in ML with SQL

- Execute, iterate, and automate ML initiatives all within BigQuery using predefined models
- Leverage external models developed in Tensorflow directly from SQL
- Export developed models for use in Vertex AI

Built-in AI/ML | Vertex AI

Why BigQuery?

- Simplify operationalizing models
- Manage MLOps in an environment designed to work with your Data Warehouse

VS

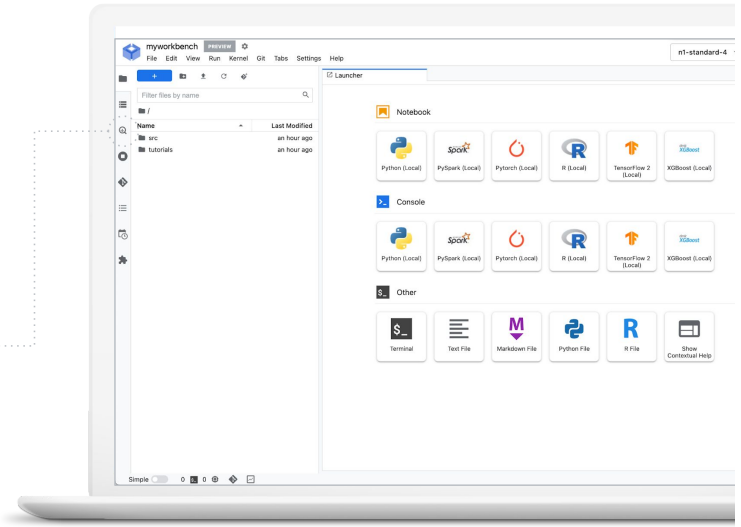
- Only use models for experiments and not operations
- Manage separate tooling for ML workloads and data integrations

Direct integration with Vertex AI Workbench

Data scientists can use Spark, Pytorch, and Tensorflow in connection with data from BigQuery with a simple integration

Complete MLOps environment

Vertex AI provides an environment for continuous training and deployment workflows with ML pipelines



Google Cloud

BI Engine for BigQuery

Powering real-time analytics

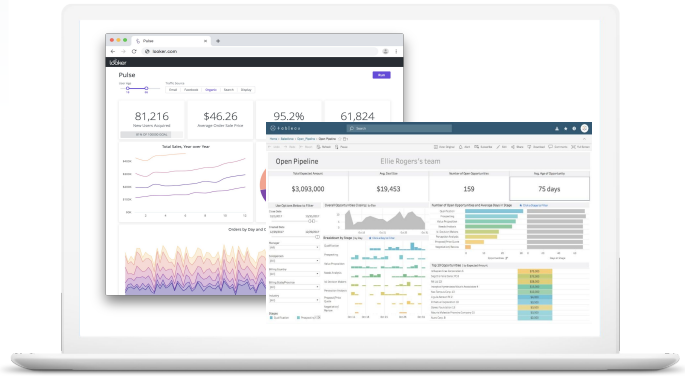
Why BigQuery?

- Accelerate SQL queries from any tool with automatic caching

VS

- Hand optimize queries or synchronize with offline BI tool caches

- Adaptive cache automatically keeps hottest data in memory
- Vectorized runtime means BigQuery queries execute in sub-second
- Use BigQuery SQL directly from Tableau, Looker, PowerBI, BigQuery UI, etc.



Data governance & security out of the box

Protecting against data exfiltration

Why BigQuery?

- A security-first approach that brings Google's security thought leadership directly to your data

VS

- Security as a secondary featureset

Fine-grained access controls through BigQuery column-level security & integration with Cloud IAM

Metadata management and data discovery in BigQuery through Dataplex

Default encrypted with data encryption in transit and at rest

Discover, classify and redact sensitive data in GCP with Cloud DLP

Shared datasets helps securely share read-only data both internally and externally at scale

Find in the documentation:

Q: What is the maximum number of tables you can create in BigQuery?

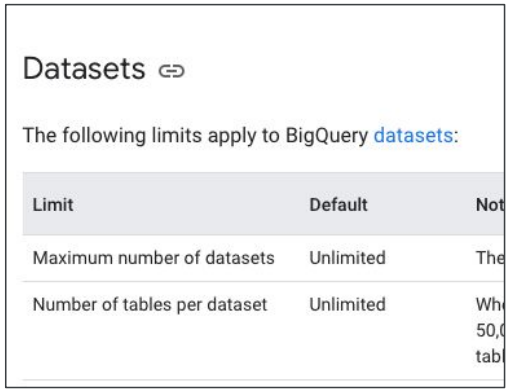
Google Cloud

Find in the documentation:

Q: What is the maximum number of tables you can create in BigQuery?

A: Unlimited

https://cloud.google.com/bigquery/quotas



Datasets

The following limits apply to BigQuery datasets:

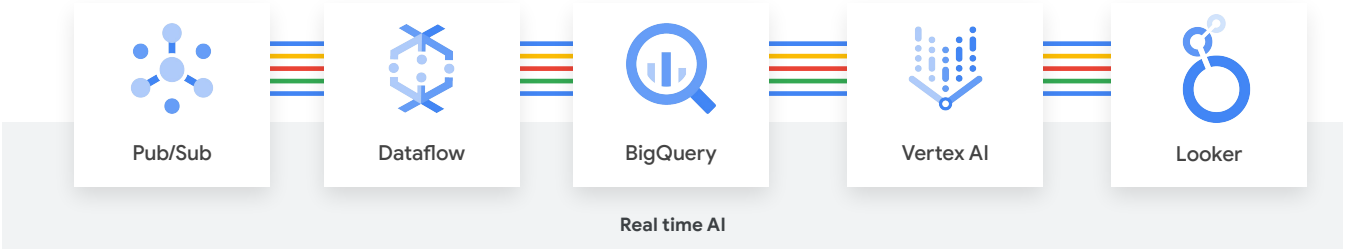
Limit	Default	Not
Maximum number of datasets	Unlimited	The
Number of tables per dataset	Unlimited	Wh 50, tab

Google Cloud

How can you get your data into BigQuery?

Google Cloud

Google's Cloud Native Golden Path for Real Time Intelligence



ML predictions

Using local and remote models with built in functions in Dataflow

Intelligent solutions

- Anomaly detection
- Object detection
- Video/vision intelligence
- Clickstream analytics
- Time-series analytics

Choice of ML frameworks

- TensorFlow
- PyTorch
- scikit-learn
- NVIDIA RAPIDS

Continuous model training

- ML CI/CD
- MLOps
- Vertex AI
- Kubeflow Pipelines
- BigQuery ML

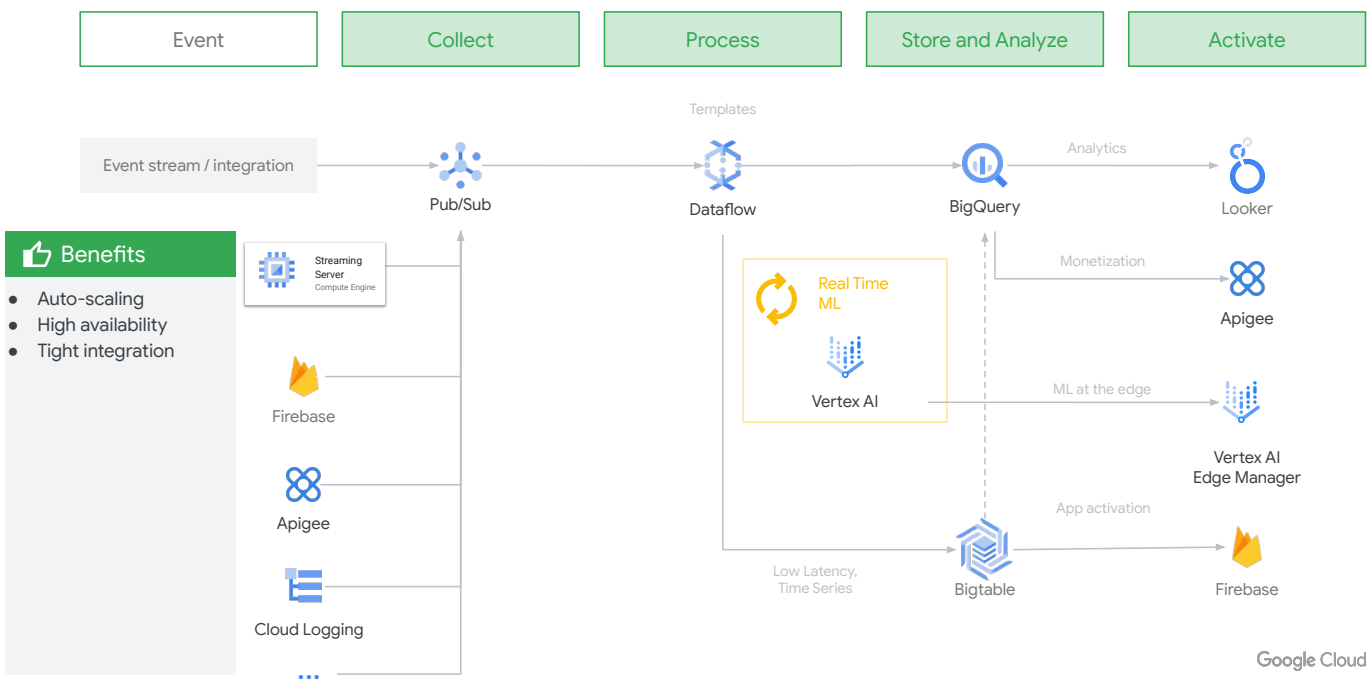
Insights activation

Downstream integrations:

- Looker / Looker Studio
- BigQuery
- BigTable
- Pub/Sub
- Apigee & others

Real Time Intelligence - High Level Architecture

Proprietary + Confidential



Pub/Sub

Simple, hyperscale, durable subscription messaging when people, applications and machines need to connect insight to everything

Pub/Sub

Messaging for **real-time analytics** and **event-driven architecture**

Scalable, durable event ingestion and delivery
Serverless, auto-scaling, and partition free; 1KB/s to 100GB/s with consistent performance; 3x3 data replication; storage for up to 31 days

Scaled Publish/Subscribe Pattern
Up to 10,000 subscriber apps per topic; independent capacity and usage-based billing for publishers and subscribers; push and pull delivery modes

Global Routing
Publish and subscribe to events regardless of geography, as simple as an API call

Deeply Integrated
Scalable analytics with Dataflow, serverless actions with Functions, power GKE autoscaling, notifications from GCS, direct to BigQuery subscriptions, all with built-in monitoring, audit logging, and compliance.

Dataflow

The backbone of of data analytics on Google Cloud, with data pipeline optimization and automation, removing the complexity of streaming intelligence

Dataflow

Highly automated, self-healing data pipeline optimization for **streaming and batch**

Ultra simplicity
Serverless, auto-provisioning, and self healing; automated performance, work balancing, unified batch and streaming under one lane

Accessible ML
Out of the box support for MLOps (TFX, Kubeflow), batch and streaming ML Inferencing (PyTorch, TensorFlow, and scikit-learn), GPU and large scale data processing

Best of OSS and optimized platform
Innovate with Apache Beam SDK; create pipelines in your language of choice; optimize workloads with built in monitoring and observability

Built to scale
Designed and built for horizontal and vertical scaling (first in industry). Optimize pipeline performance at any scale; maximize utilization to save costs

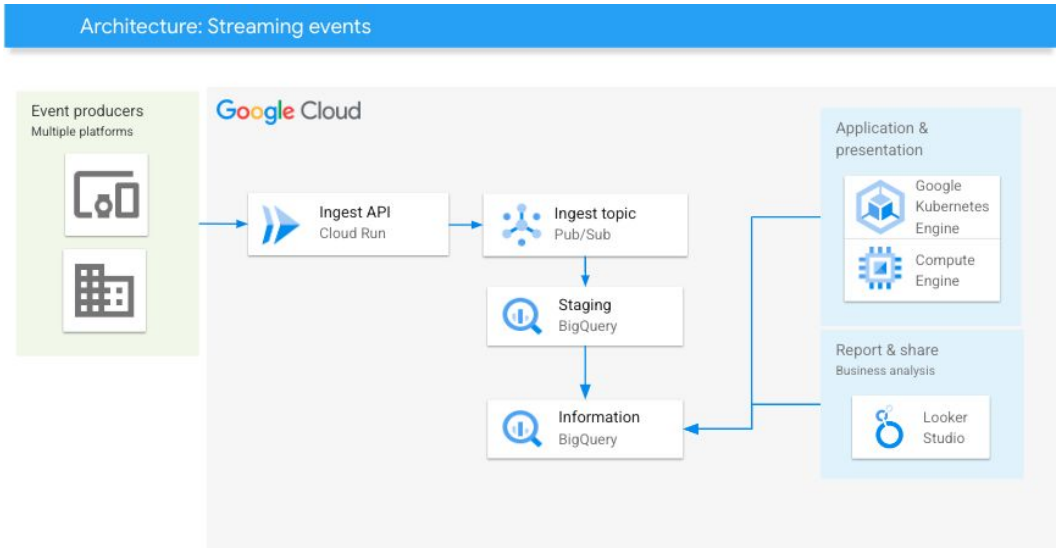
Read more: https://cloud.google.com/blog/topics/developers-practitioners/dataflow-backbone-data-analytics

Stream events in real-time for data analytics on Google Cloud

This solution explores an example architecture pattern that ingests, processes, and analyzes a large number of events concurrently from many different sources. The processing happens as events unfold, enabling you to respond and make decisions in real-time.

Pub/Sub acts a pipeline for ingestion of real-time events. Dataflow performs the data transformation. That data is loaded into the BigQuery analytics engine.

Looker Studio or Looker can be used for creating dashboards and visualizing the data.



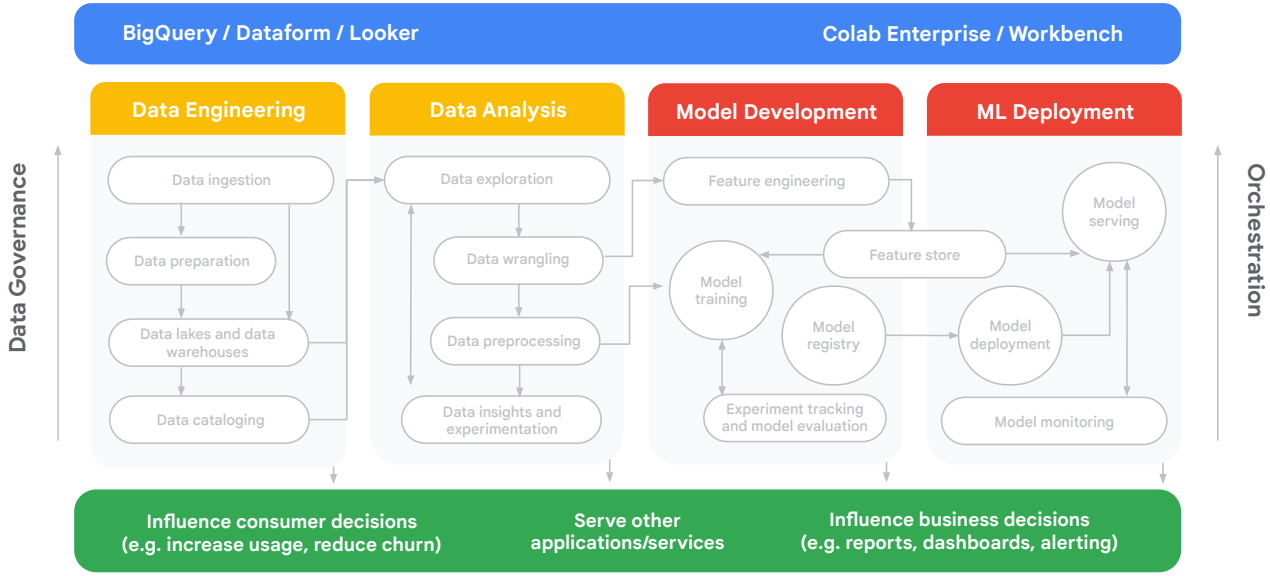
Try it yourself

Demo :
goo.gle/loff-demo24-StreamAnalytics

More solutions:
goo.gle/loff-demo24-click-to-deploy

Data and AI platform

Build, experiment, deploy, and manage ML models.



Get started with AI

Google Cloud

What is Artificial Intelligence?

Google Cloud

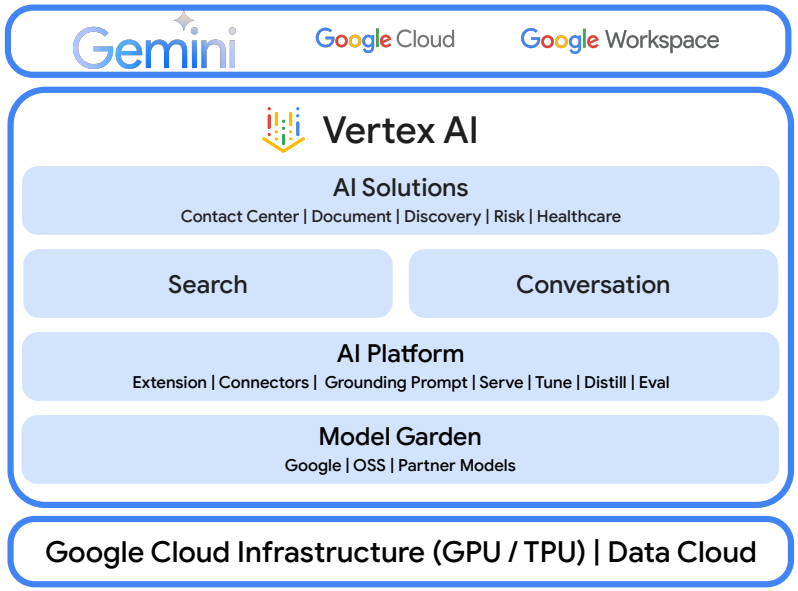
What is Artificial Intelligence?

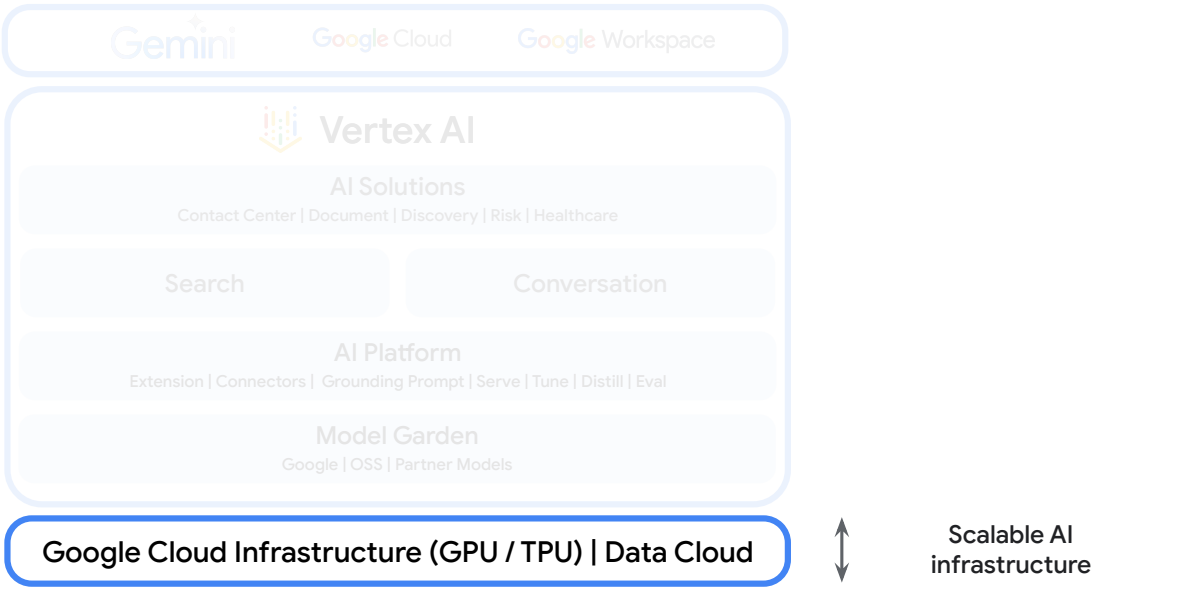
Artificial intelligence (AI) is a set of technologies that enable computers to perform a variety of advanced functions, including the ability to see, understand and translate spoken and written language, analyze data, make recommendations, and more.

Try me!

Google Cloud

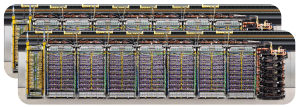
Let’s take a look at how this manifests in our portfolio




Gemini
Google Cloud
Google Workspace
Vertex AI
AI Solutions
Contact Center | Document | Discovery | Risk | Healthcare
Search
Conversation
AI Platform
Extension | Connectors | Grounding Prompt | Serve | Tune | Distill | Eval
Model Garden
Google | OSS | Partner Models
Google Cloud Infrastructure (GPU / TPU) | Data Cloud
Scalable AI infrastructure

Purpose-built infrastructure for Generative AI

Tensor Processing Unit: Designed by Google for AI at Scale



Cloud TPU v5e
up to 2.7x Inference Perf/$ vs v4

Cloud TPU v5p
up to 2.8x LLM training vs v4

Cost Efficient & Versatile
(Training & Inference)

Powerful & Flexible
(full range of AI models)

NVIDIA GPUs: Latest NVIDIA GPUs on Google Cloud

NVIDIA

G2 GPU VM
2-4x performance improvement vs T4

Powered by
NVIDIA L4 GPU

A3 GPU VM
3x training improvement vs A2 VM

Powered by
NVIDIA H100 GPU

- Provide a **wide variety of hardware** options
- **Scale** AI models exponentially
- Leverage our fully-managed AI platform optimized for **efficiency**
- Build with an **open source software ecosystem**

Cloud TPU v5e

GA

Highly efficient, versatile, and scalable AI supercomputer

Cost Efficient

Up to 2.3X Training Perf/$
(Vs TPU v4)

Up to 2.7X Inference Perf/$
(Vs TPU v4)

Versatile

Training & Inference
(bfloat 16, Int8, Int4)

8 VM Shapes
(To fit diverse model sizes)

Leading AI Frameworks
(PyTorch, JAX, TensorFlow)

Scalable

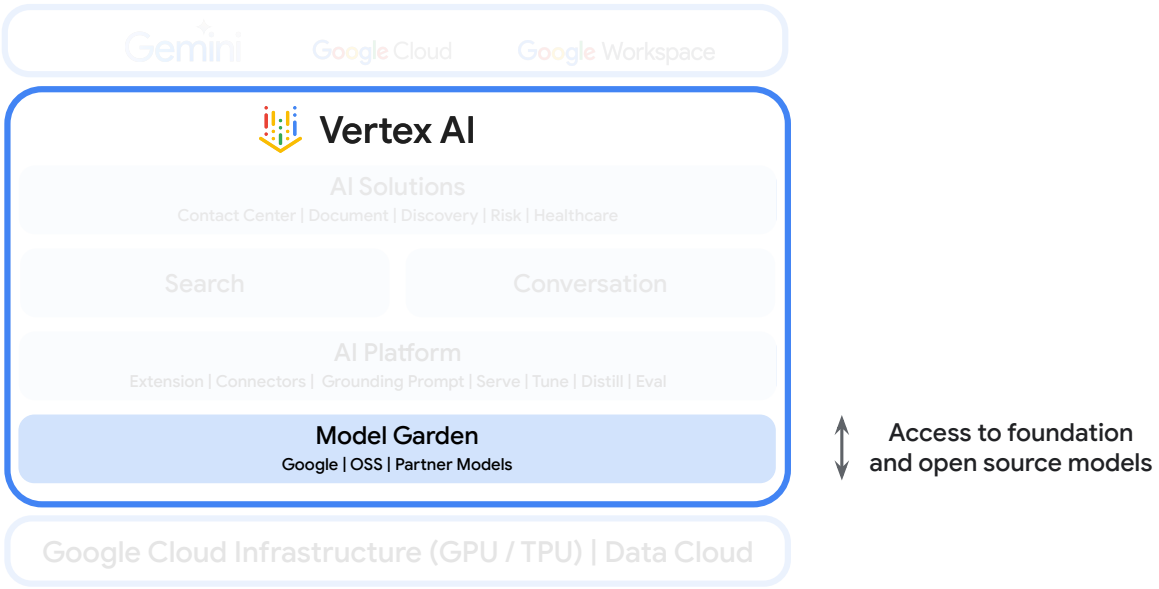
Scale to 10s Ks of Chips
(Multislice Training)

Interchip Interconnect & DCN
(Same XLA programming model)

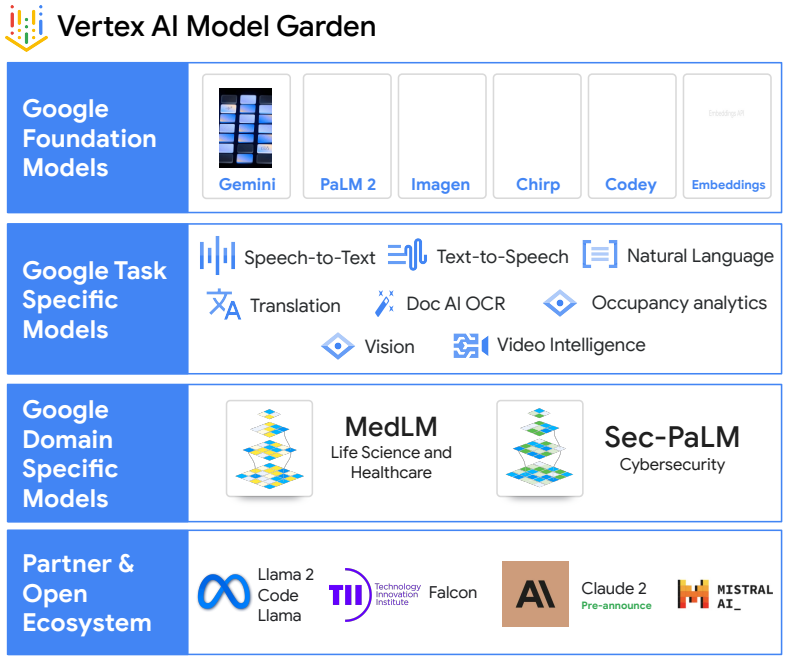
Horizontal Scaling in GKE
(Auto-provisioning & Scaling)

Training: Performance measured by Google, seq-len=2048 for GPT-3 175 billion parameter model implemented using paxml, seq-len = 2048 for 32 Billion parameter decoder model using MaxText; Relative performance using price of TPU v4: $3.22/chip/hour and TPU v5e: $1.2/chip/hour
Inference: All results measured by Google and normalized to single-chip throughput. Precision: Llama 2 7B, 13B, 70B, GPT-J 6B: int8; GPT-J 175B, Stable Diffusion 2.1: bf16.

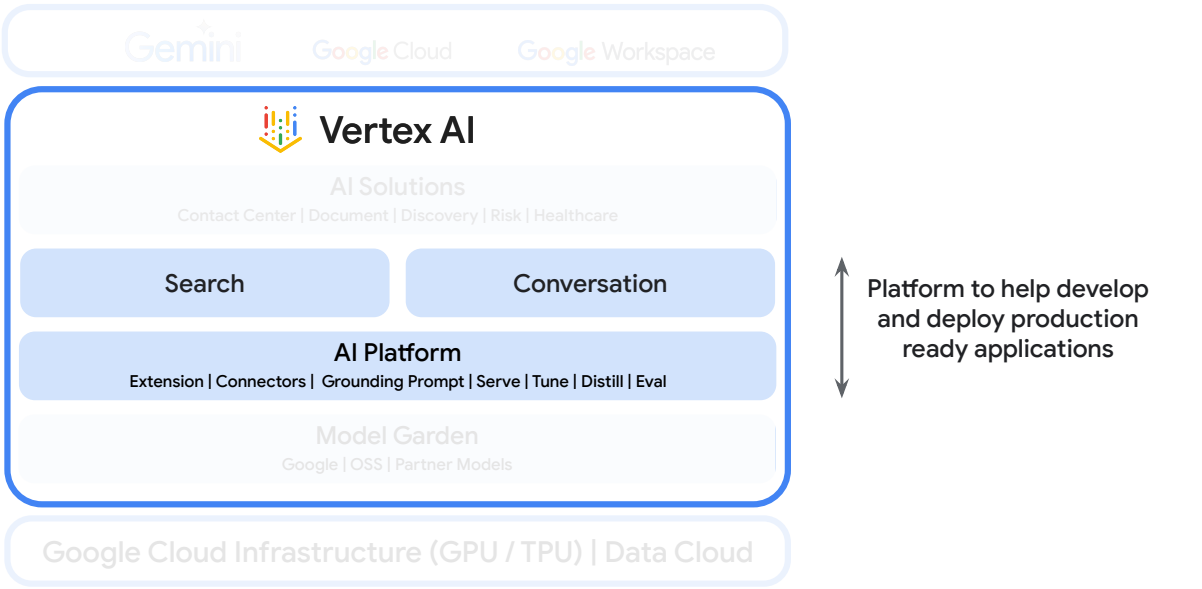
Google Cloud


Gemini
Google Cloud
Google Workspace
Vertex AI
AI Solutions
Contact Center | Document | Discovery | Risk | Healthcare
Search
Conversation
AI Platform
Extension | Connectors | Grounding Prompt | Serve | Tune | Distill | Eval
Model Garden
Google | OSS | Partner Models
Google Cloud Infrastructure (GPU / TPU) | Data Cloud
Access to foundation
and open source models

130+ enterprise-ready foundation models in Vertex AI Model Garden

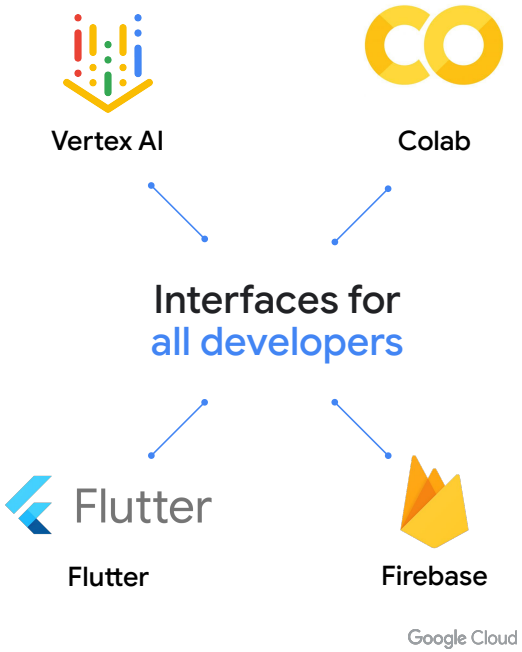


- **Choice and flexibility** with Google, open source, and third-party foundation models
- **Multiple modalities** to match every use case
- **Multiple model sizes** to match cost and efficacy needs
- **Domain-specific models** for specialized industries
- Enterprise ready with **safety, security, and responsibility**
- Decrease time to value with **fully integrated platform**

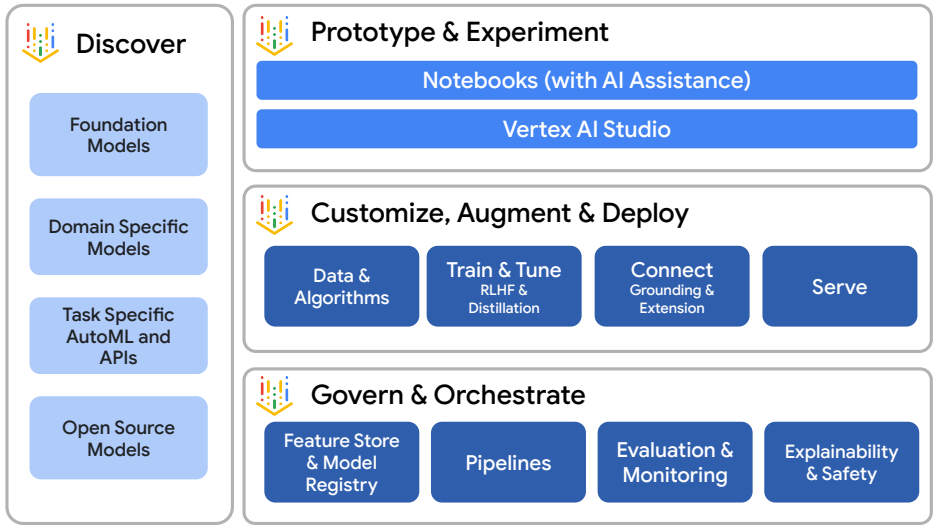

Gemini
Google Cloud
Google Workspace
Vertex AI
AI Solutions
Contact Center | Document | Discovery | Risk | Healthcare
Search
Conversation
AI Platform
Extension | Connectors | Grounding Prompt | Serve | Tune | Distill | Eval
Model Garden
Google | OSS | Partner Models
Google Cloud Infrastructure (GPU / TPU) | Data Cloud
Platform to help develop and deploy production ready applications

Vertex AI is built for developers

- Extensive **quick start library** with code samples and jumpstarts for **developers of all levels** and ecosystems
- **No cost developer labs** and training resources across Vertex products at Cloud Skills Boost
- **Robust integrations** with popular third party developer tools like **Lang Chain, LlamaIndex, Pinecone, and Weaviate.**
- **Packages and extensions** to natively support Google Cloud foundation models in Google app developer frameworks like **Firebase and Flutter**.

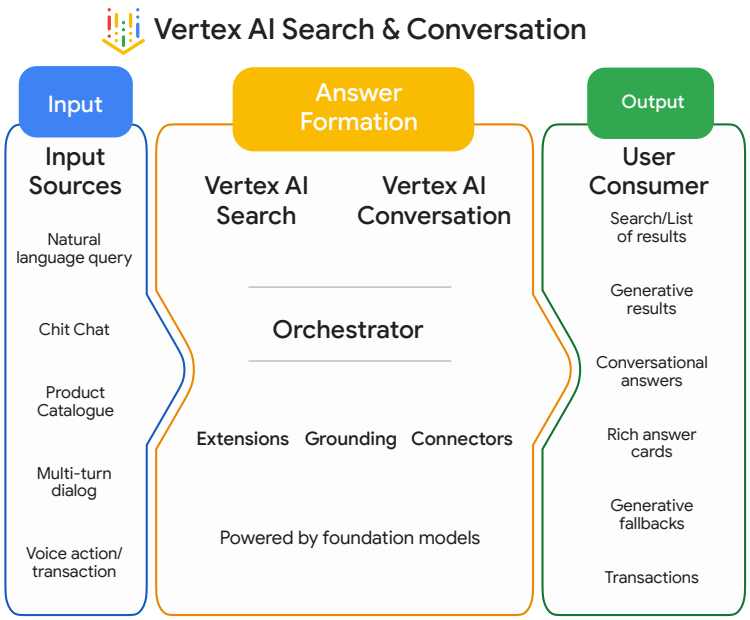


The best of predictive and generative AI in one platform

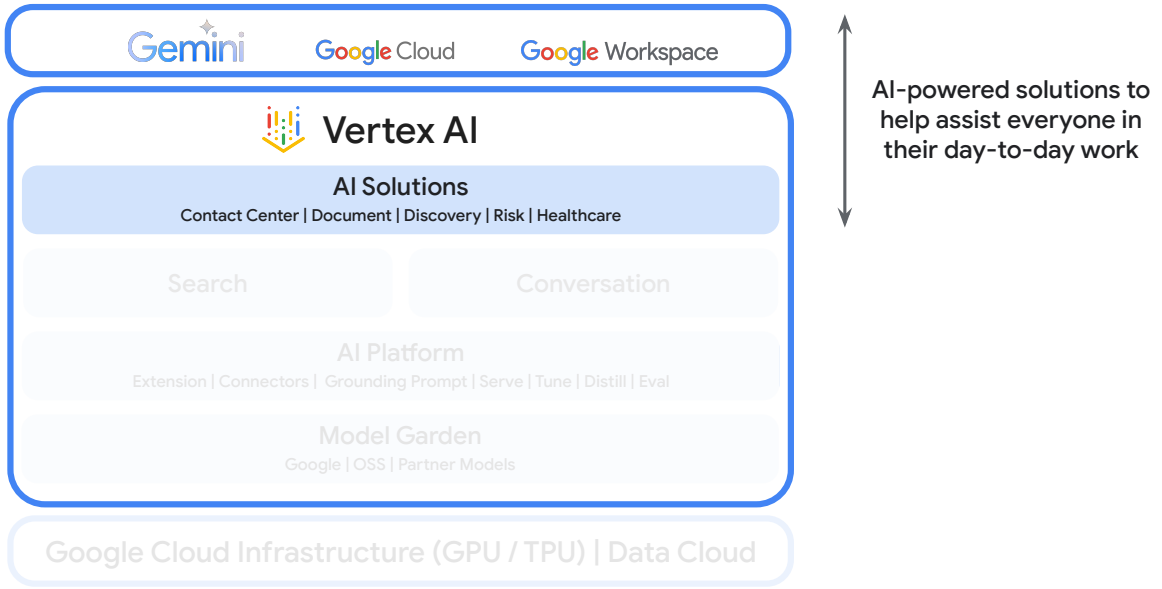


- Unified interface for **data analysis and AI development** to kickstart AI-powered projects across various tasks & modalities
- **Augmentation tooling** to differentiate yourself using your enterprise data
- Consistent workflows with minimal data movement **simplify data governance** for every enterprise
- Unification of **predictive and generative** through MLOps

Low code developer platform for Gen AI-powered Search and Conversation experiences



- Empower developers to help build **personalized** conversational experiences
- Ground foundation models with **fresh and factual data** to reduce hallucinations
- Connect foundation models to **real-time actions and real-time data** to scale helpfulness
- Leverage years of **information retrieval and deep retrieval** experience from Google Search
- Simplify known use cases with **pre-built tasks, playbooks, and orchestrators**


Gemini
Google Cloud
Google Workspace
Vertex AI
AI Solutions
Contact Center | Document | Discovery | Risk | Healthcare
Search
Conversation
AI Platform
Extension | Connectors | Grounding Prompt | Serve | Tune | Distill | Eval
Model Garden
Google | OSS | Partner Models
Google Cloud Infrastructure (GPU / TPU) | Data Cloud
AI-powered solutions to help assist everyone in their day-to-day work

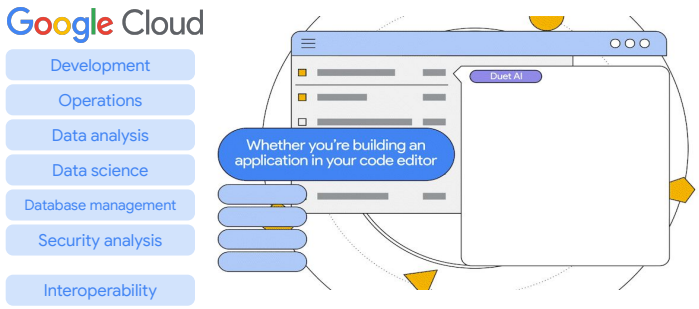
Always on AI collaborator for everyone, Gemini for Workspace

Helps you write in Gmail and Docs

Helps you organize in Sheets

Helps you visualize in Slides

Helps you connect in Meet

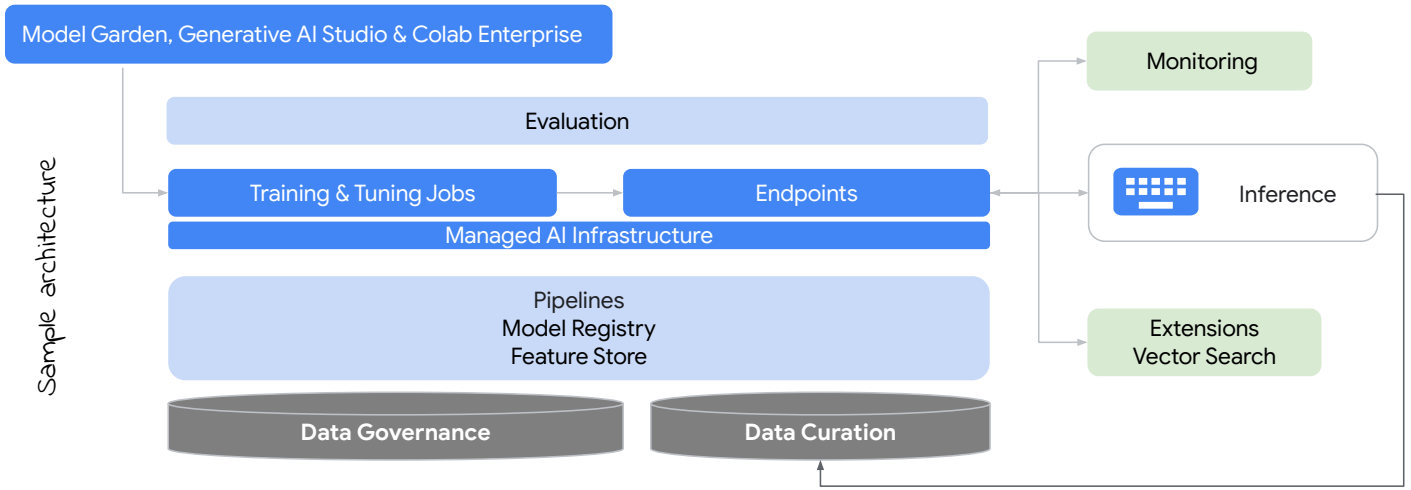


- **Democratize generative AI** for all users through **expert assistance** through Google Cloud
- Help increase employee productivity in Workspace by streamlining **creation, connection, and collaboration**
- Help drive **developer efficiency** through in the developer IDE, Cloud Console, databases, and security products
- Allow developers to focus on the **most value-add aspects** of their job

What does MLops look like on Vertex AI?

MLOps on Vertex AI

Same Platform, Upgraded Capabilities for predictive AND generative AI



Colab Enterprise on Vertex AI

Colab Enterprise combines the ease of use of Google Colab notebooks with the enterprise-level security and compliance capabilities of Google Cloud

GA

Use Cases: Data science, data analysis, data engineering, ML engineering

Collaboration & Productivity

IAM based notebook sharing

Automatic Versioning

Commenting (coming soon!)

Co-editing (coming soon!)

Generative AI powered code completion and generation

Zero-Config & Flexible Compute

Provides both zero-config compute options, as well as access to a wide range of machine-shapes and compute

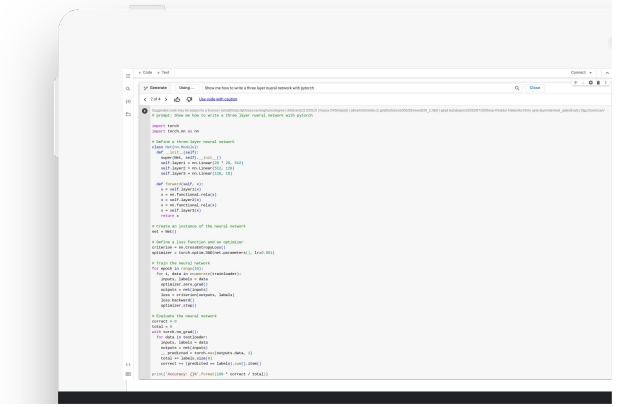
Enterprise Ready

Will support a wide range of security and management capabilities including:

- VPC-SC
- CMEK
- Regionalization
- Cloud Monitoring
- Cloud Logging

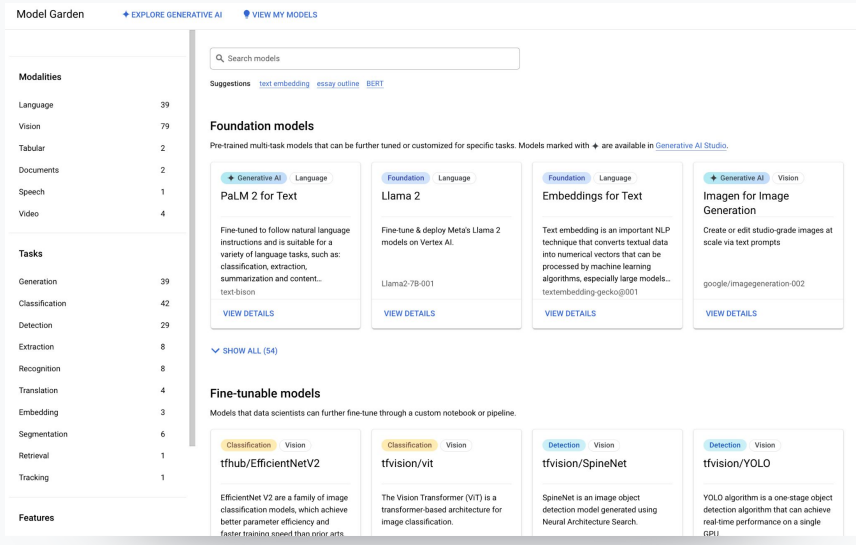
Available across Google Cloud

Available in BigQuery and Vertex AI (Dataproc coming soon), making it easy to work across data and AI workloads



Model Garden

Access, customize, and experiment with Google, OSS, and third party models and APIs



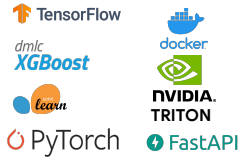
Deploy & Serve pre-trained and custom models

Discover and try pre-trained and open foundation AI models available on **Vertex Model Garden**

Carefully author an input prompt with clear, concise and informative instructions on **Vertex Generative AI Studio**

Deploy to fully managed endpoints with **Vertex Predictions** supporting autoscaling, private endpoints, and wide selection of CPUs & GPUs

- Model co-hosting
- Custom prediction routines

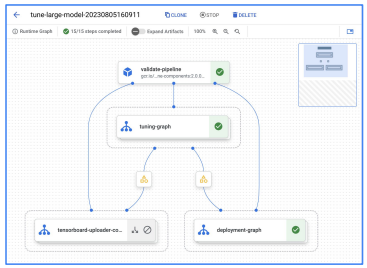


Powered by AI Infrastructure for your generative AI needs

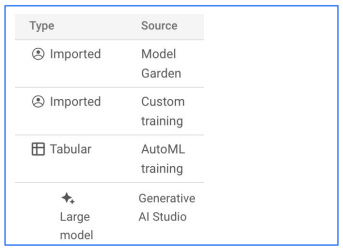
NVIDIA GPUs A100, L4 and H100 | TPU v3, v4 and v5e

Managing artifacts

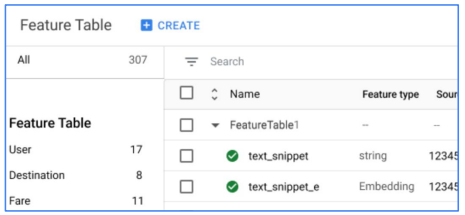
Orchestrate & manage training & tuning jobs via **Vertex AI Pipelines** to trace lineage from dataset to model



Manage predictive & generative AI models in **Vertex AI Model Registry** with access to evaluation metrics and one-click deployment



Store, manage, and serve features including embeddings in **Vertex AI Feature Store** to support predictive & generative AI needs



Evaluation services for predictive AI

Evaluate

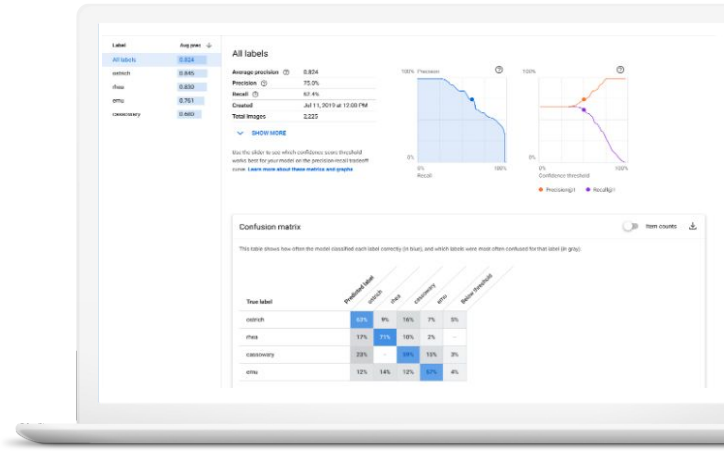
Iteratively and independently run model evaluations on new datasets at scale

Compare

Visualize and compare different model evaluations to identify the best model for production deployment

Assess

Assess the performance of models on different slices and evaluated annotations



Evaluation services for generative AI

1

Automatic Metrics

rougeLSum	0.1345
bleu	0.763

Assess the performance of a model with task-specific metrics computed based on reference data

- Fast and efficient
- Standard method used in academia and many open benchmarks

2

AutoSxS

model_a_win_rate	0.35
model_b_win_rate	🏆 0.65

Compare the performance of 2 models with an arbiter model

- Consistent performance with human evaluation, while being faster, cheaper, available on demand

3

Safety Bias

Safety attribute: Toxicity	
GenderID: female	0.89
GenderID: male	0.12
GenderID: transgender	0.02
GenderID: non-binary	0.67

Understand if the safety profile of your model is biased against a certain identity group.

Model Monitoring

Monitor & Explain

- Model monitoring for skew & drift detection
- Feature-based explability
- Example-based explainability

Safety Scores

Score “harmful categories” and topics that may be considered sensitive.


```
{
"predictions": [
    {
      "safetyAttributes": {
        "categories": [
          "Hate",
          "Toxic",
          "Violent",
          "Sexual",
          "Insult",
          "Obscene",
          "Death, Harm & Tragedy",
          "Firearms & Weapons",
          "Public Safety",
          "Health",
          "Religion & Belief",
          "Drugs",
          "War & Conflict",
          "Politics",
          "Finance",
          "Legal"
        ],
        "scores": [
          0.7,
          0.4,
          0.6,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0
        ],
        "blocked": false
      },
      "content": "<>"
    }
```


Recitation Checks

Help ensure that our models do not replicate existing content at length

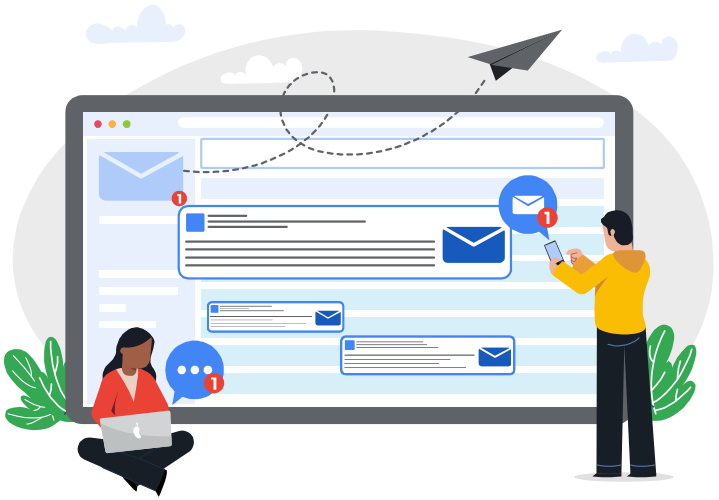
Security with Google Cloud AI

Data governance during generative AI model usage

Google processes prompts to provide the service.

Prompts input to the foundation model to generate a response, are encrypted in transit.

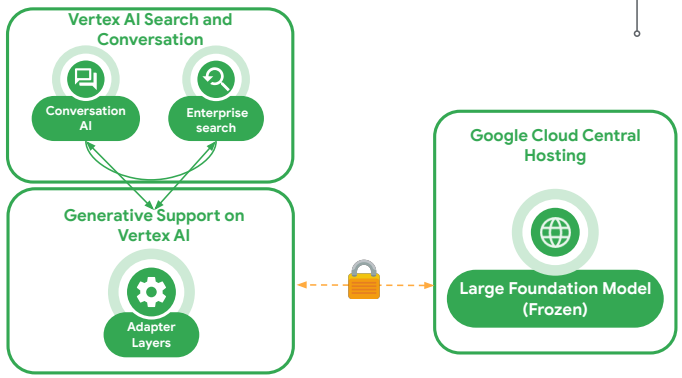
Google does not use prompt data to train its models without the express consent of its customers.



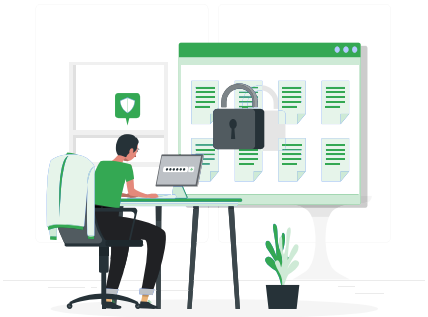
Google Cloud

Security during Model Tuning

- Parameter Efficient Fine Tuning (PEFT)
- Customer specific adapter weights
- Foundational model remains frozen during inference



- Input data is stored securely
- Adapter weights are stored securely
- Customer can delete adapter weights at any time
- Customer Data will not be logged to train foundation models by default



Security Controls for Generative AI

VPC Security Controls: Help protect against accidental or targeted action by external entities or insider entities, to minimize unwarranted data exfiltration risks.

Customer Managed Encryption Keys: Control the keys that protect your data at rest. Support for External Key Management

Access Transparency: Near real-time logs offer insight when Google Cloud administrators access your content - including access justification.

Technical safeguard: Safety filters


```
{
"predictions": [
    {
      "safetyAttributes": {
        "categories": [
          "Hate",
          "Toxic",
          "Violent",
          "Sexual",
          "Insult",
          "Obscene",
          "Death, Harm & Tragedy",
          "Firearms & Weapons",
          "Public Safety",
          "Health",
          "Religion & Belief",
          "Drugs",
          "War & Conflict",
          "Politics",
          "Finance",
          "Legal"
        ],
        "scores": [
          0.7,
          0.4,
          0.6,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0,
          0.0
        ],
        "blocked": false
      },
      "content": "<>"
}
```


Safety attributes label “harmful categories” and topics that may be considered sensitive.

Each safety attribute has an associated confidence score in [0.0, 1.0], rounded to one decimal place, **reflecting the likelihood of the input or response belonging to a given category**.

You can test Google’s safety filters and define confidence thresholds that are right for your business and take comprehensive measures to detect content that violates Google’s usage policies or terms of service.

Recitation checks

Recitation checks help ensure that our models do not replicate existing content at length

We've designed our systems to limit the chances of replicating existing content at length and we will continue to improve how these systems function.

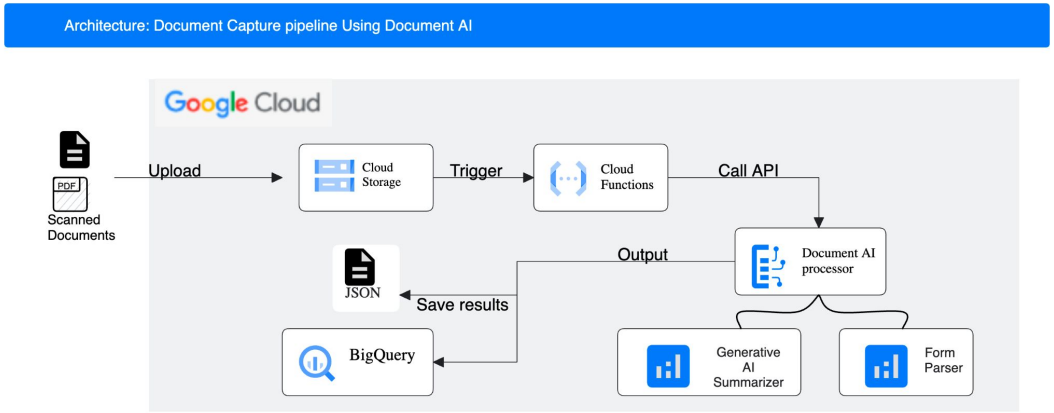
Generally, if our API does directly quote at length from a page on the web that was part of our pre-training dataset, it cites that page in our output from the model, or if over a certain length, it will block the output.

Extract data from your documents using Generative AI

Automatically extract structured data (e.g., from forms, invoices) using Document AI's form processor.

The solution leverages genAI to produce concise summaries of the processed documents, highlighting key information for enhanced understanding.

Extracted data and summaries are stored in BigQuery, enabling powerful analysis, visualization, and machine learning applications.



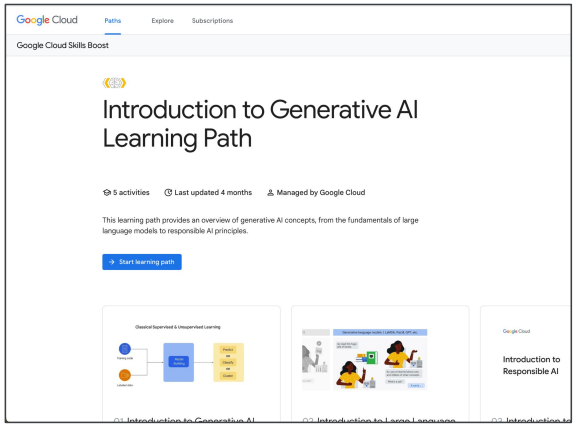
Try it yourself

Demo :
goo.gle/loff-demo24-DocAIForm

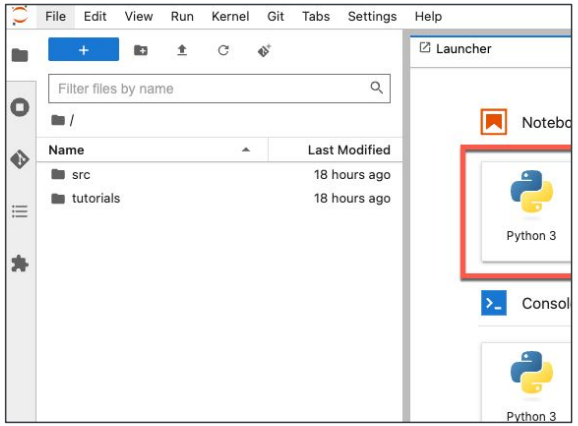
More solutions:
goo.gle/loff-demo24-click-to-deploy

Try it yourself

[Learn more about Generative AI]



[Train a model using Vertex AI and Python SDK]



Google Cloud

What's next?

Google Cloud

What is Google Cloud?

It’s how you can...

Session 1
Host your first application

Session 2
Store and manage your data

Session 3
Analyze your data

Session 4
Get started with AI

Thank you

Check out more ways to learn
in the description below

Graveyard

Limitless data

Limitless data scale

275EB

analyzed across BigQuery
in December 2021

110+TB

data analyzed per second

> “...BigQuery continued to **scale** in storage, compute, concurrence, ingest and reliability as we added **more and more** users, traffic, and data.”

Nikhil Mishra
Sr. Director of Engineering, Verizon Media

verizon

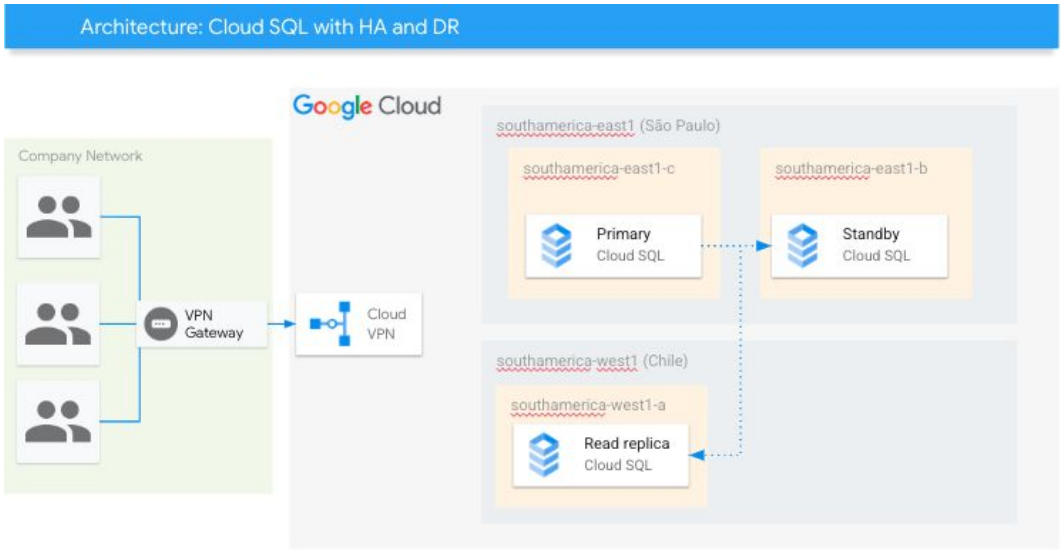
Host a Highly Available SQL database on Google Cloud

Cloud SQL instance with high-availability and cross-region replica.

Leverages regional and zonal failovers as well as read replicas offered by Cloud SQL

Uses Cloud VPN to connect company network to Google Cloud

Suitable for a wide range of applications and industries, including e-commerce platforms, financial systems, customer relationship management (CRM) tools, and more.



Try it yourself

Demo :
goo.gle/loff-demo24-CloudSQLMR

More solutions:
goo.gle/loff-demo24-click-to-deploy