Rad AI

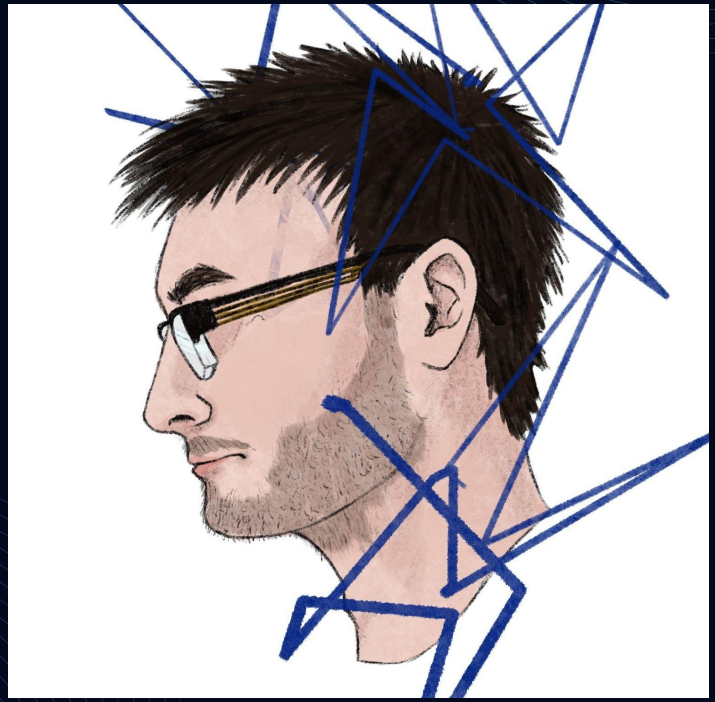
GCP Services Overview

Fletcher Easton, Reporting

November 4th, 2024

Introduction

- Fletcher Easton
- Full-Stack Engineer on Reporting
- Joined September 26, 2023
- Previously at ROSALIND
  - Large-scale bioinformatics
  - Created and managed our FedRAMP'd platform for COVID-19 assay analysis (before GCP's guide)



Agenda

Run Code

- Compute Engine
- GKE
- Cloud Run
  - Services
  - Functions
  - Jobs

Store Data

- Cloud Storage
- Firestore
- Cloud SQL
- AlloyDB
- Cloud Spanner

Misc.

- Artifact Registry
- BigQuery
- Cloud Build

Run Code – Compute Engine

What is it?

- Managed Virtual Machines
- Control the OS, vCPUs, RAM, Storage, etc.
- Run whatever you want

What are the costs*?

- Fixed cost for having the VM on
  - # of vCPUs
  - Amount of RAM
- Fixed cost for having the VM
  - Storage capacity and type
- Optional costs
  - “Premium” images
  - Sole-tenancy

Run Code – Compute Engine

Machine configuration

General purpose | ✓ Compute optimized | Memory optimized | Storage optimized | GPUs

Machine types for performance-intensive workloads, with highest performance per core

	Series	Description	vCPUs	Memory
○	H3	High performance computing workloads	88	352 GB
◉	C2	Ultra-high performance, compute-intensive workloads	4 - 60	16 - 240 GB
○	C2D	Ultra-high performance, compute-intensive workloads	2 - 112	4 - 896 GB

Machine type

Choose a machine type with preset amounts of vCPUs and memory that suit most workloads.

c2-standard-8 (8 vCPU, 4 core, 32 GB memory)

vCPU
8 (4 cores)

Memory
32 GB

Monthly estimate

$260.89

That's about $0.36 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
8 vCPU + 32 GB memory	$304.86
100 GB SSD persistent disk	$17.00
Use discount	-$60.97
Total	**$260.89**

Compute Engine pricing

LESS

Run Code – Compute Engine

Machine configuration

General purpose | Compute optimized | Memory optimized | Storage optimized | ✓ GPUs

Graphics processing units (GPUs) accelerate specific workloads on your instances such as machine learning and data processing. Learn More

GPU type: NVIDIA H100 80GB

Number of GPUs: 8

☐ Enable Virtual Workstation (NVIDIA GRID)

	Series	Description	vCPUs	Memory	Platfor
◉	A3	High-performance AI/ML training and HPC applications	208	1,872 GB	NVIDIA

Machine type

Choose a machine type with preset amounts of vCPUs and memory that suit most workloads.

a3-highgpu-8g (208 vCPU, 104 core, 1,872 GB memory)

vCPU
208 (104 cores)

Memory
1,872 GB

Monthly estimate

$65,873.10

That's about $90.24 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
208 vCPU + 1,872 GB memory	$6,905.84
8 NVIDIA H100 80GB	$57,211.86
6,000 GiB Local SSD disks	$480.00
10000 GB Hyperdisk Balanced	$800.00
80000 provisioned IOPS	$385.00
2400 provisioned throughput	$90.40
Total	**$65,873.10**

Compute Engine pricing

LESS

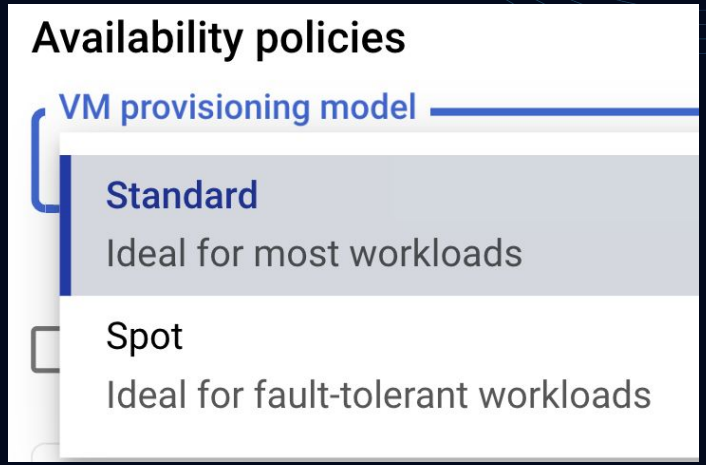
Run Code – Compute Engine

Standard VMs

- VM runs when you choose

Spot VMs

- Compute Engine “excess capacity”
- 60-90% discount for most configurations
- Great for fault-tolerant workloads
- Use in a managed instance group



Run Code – GKE

What is it?

- Clusters > Node Pools > Nodes > Pods
- Manages Compute Engine instances for your workloads
- Pods run workloads
- Pods run on Nodes
- Nodes run in Node Pools
- Node Pools specify Node config

What are the costs*?

- $0.10 / cluster / hour
- Costs for Compute Engine

Run Code – Cloud Run Services

What is it?

- Manages containerized web applications
  - Provide a container that listens to `$PORT`
  - Cloud Run handles the rest
- Based on knative
- Fast autoscaling
- Easy traffic management
- Easy rollouts/rollbacks
- Low operational overhead

What are the costs*?

- Per-request resource-based billing

Run Code – Cloud Run Services

- Sidecar
- Built-in metrics
- Secret access
- Service IAM Identity
- Rollbacks, rollouts, and traffic management
- GPU access (public preview)
- VPC networking
- Simple Cloud SQL connection
- Health checks
- HTTP/2
- WebSockets
- gRPC

Run Code – Cloud Run Functions

What is it?

- Equivalent to AWS Lambda
- Cloud Run Services, but...
  - Provide code, specify runtime
  - Respond to events

What types of events?

- HTTP requests
- Eventarc
- Pub/Sub
- Cloud Storage
- Cloud Firestore
- Cloud Audit Logs

Run Code – Cloud Run Services/Functions

Active when serving requests

- $0.0000240 / vCPU-second
  - First 50 hours a month free
- $0.0000025 / GiB-second
  - First 100 hours a month free
- $0.40 / million requests
  - First 2 million requests a month free

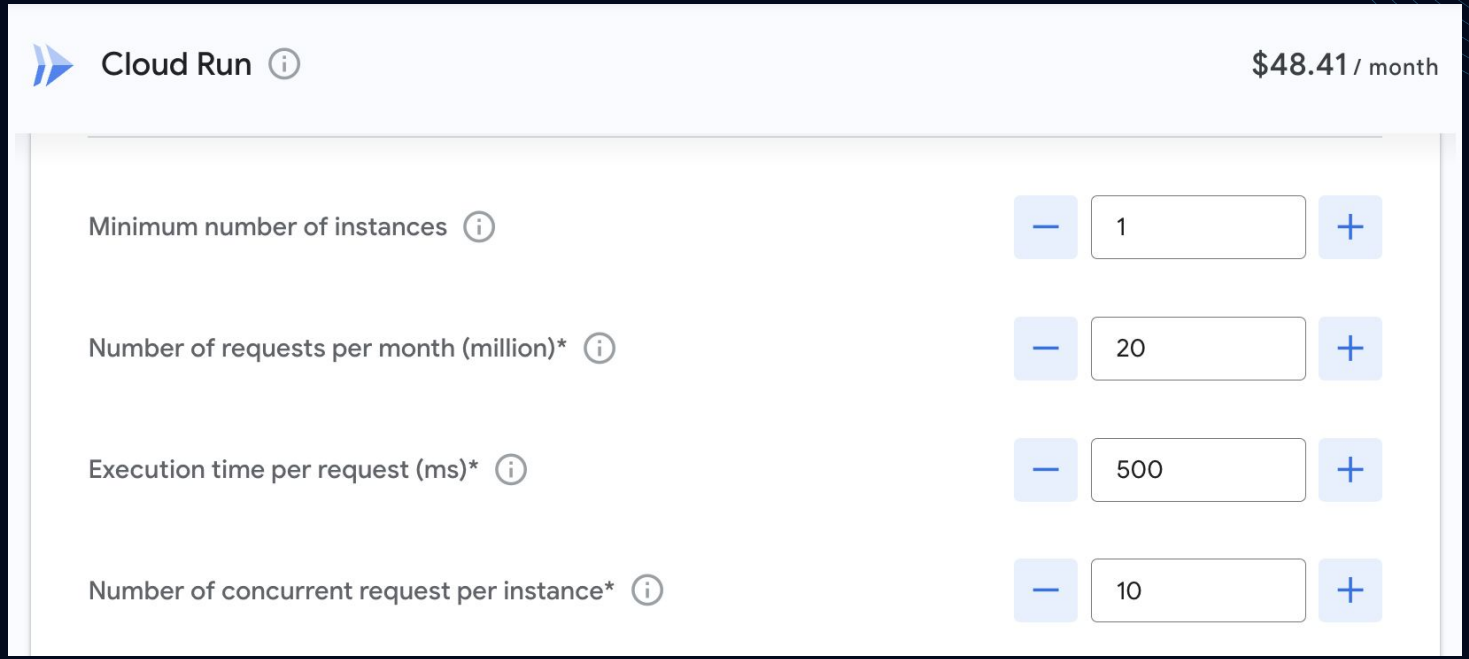
Always active

- $0.0000180 / vCPU-second
  - First 66.67 hours a month free
- $0.0000020 / GiB-second
  - First 125 hours a month free

Billed at 100ms increments

Run Code – Cloud Run Services/Functions

Pricing Calculator



Run Code – Cloud Run Jobs

What is it?

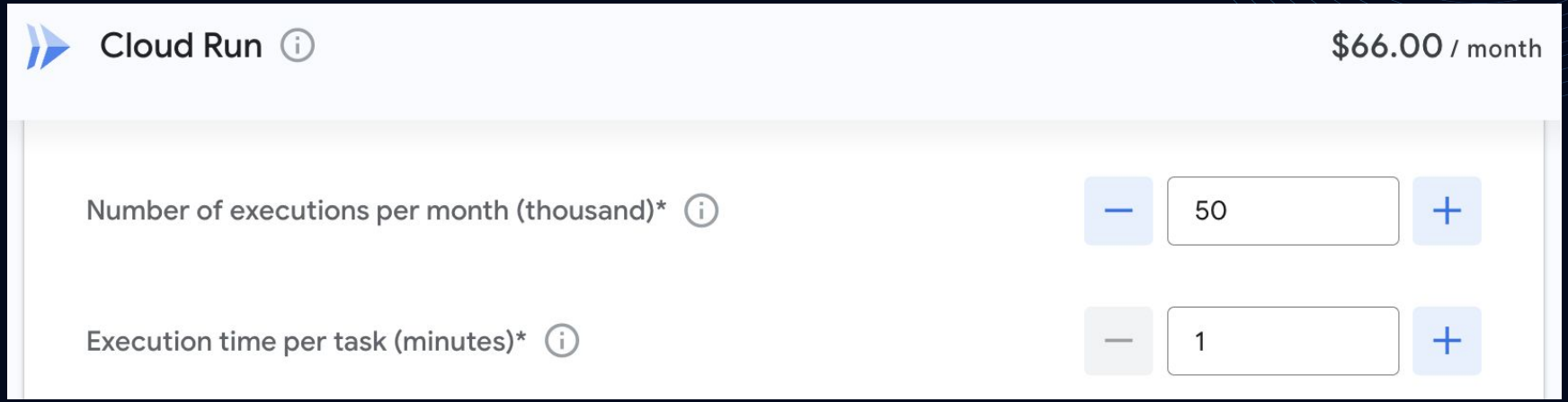
- Manages containerized data-processing applications
- Jobs can be executed...
  - On a schedule using Cloud Scheduler
  - Programmatically with arguments

What are the costs?

- Same pricing as Cloud Run Services
- Minimum billing period of 1 minute
  - Prefer batch processing when possible

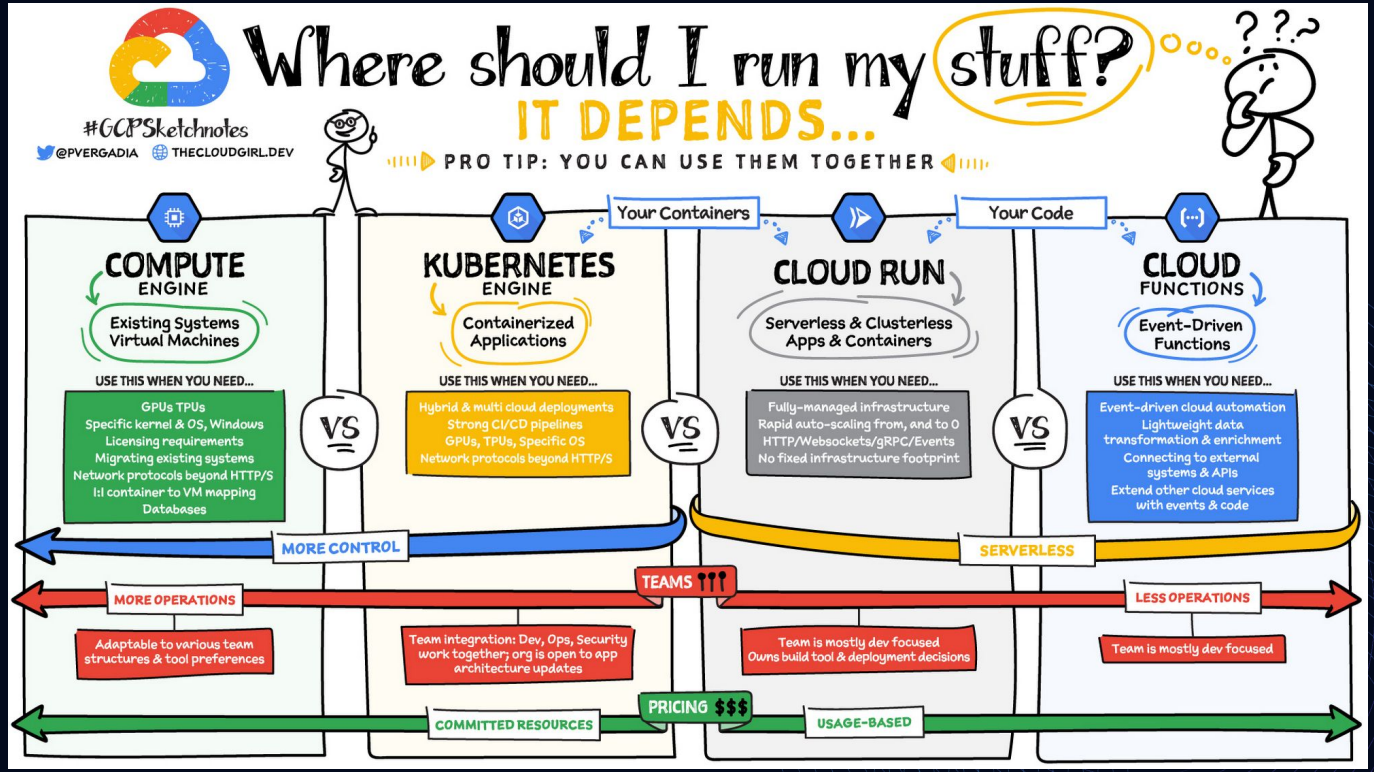
Run Code – Cloud Run Jobs

Pricing Calculator

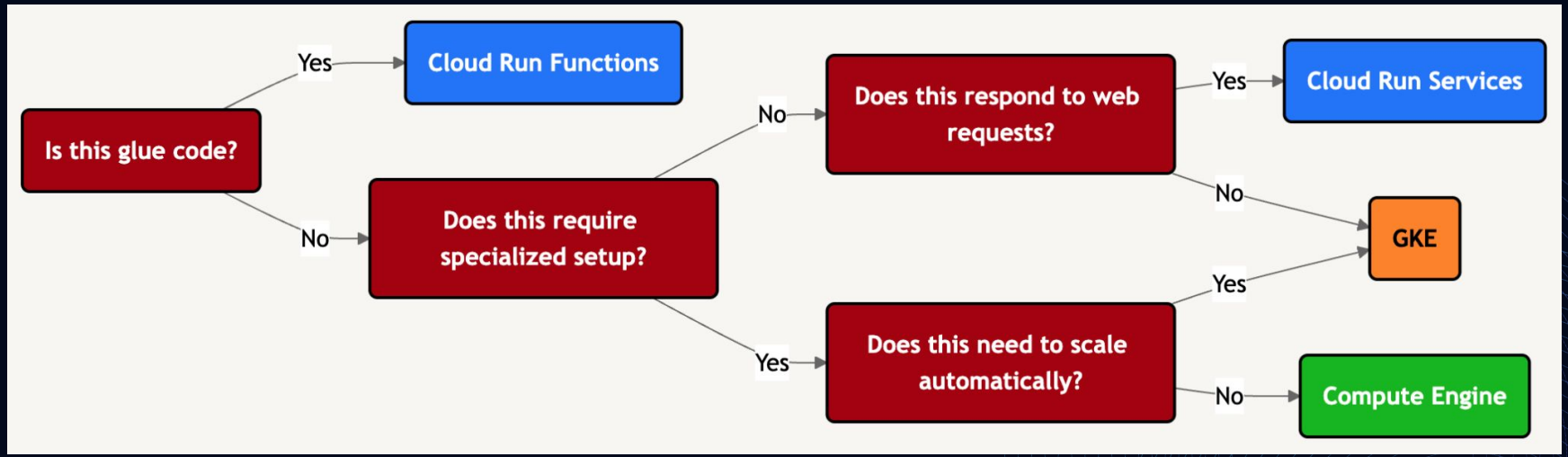


Where should I run my stuff?

GCP Blog



Where should I run my stuff?



Agenda

~~Run Code~~

- Compute Engine
- GKE
- Cloud Run
  - Services
  - Functions
  - Jobs

Store Data

- Cloud Storage
- Firestore
- Cloud SQL
- AlloyDB
- Cloud Spanner

Misc.

- Artifact Registry
- BigQuery
- Cloud Build

Store Data – Cloud Storage

What is it?

- Managed blob storage
- Equivalent to AWS S3
- 99.999% uptime
- Storage classes that determine cost and access
- Misc. features
  - Autoclass
  - Object versioning

What are the costs?

- Storage billing by class
  - `Standard: $0.0200 / GB / month`
  - `Nearline: $0.0100 / GB / month`
  - `Coldline: $0.0040 / GB / month`
  - `Archive:  $0.0012 / GB / month`
- Operations and data access by class

Store Data – Firestore

What is it?

- Managed, serverless NoSQL document database
- Allows access to the database from the client with live updates and offline mode

What are the costs*?

- Storage: $0.15 / GB / month
- Writes: $0.09 / 100k
- Reads: $0.03 / 100k
- Deletes: $0.01 / 100k

Store Data – Cloud SQL

What is it?

- Managed SQL database
  - Postgres
  - ~~MySQL~~
  - ~~SQL Server~~
- Automated backups, replicas, maintenance, etc.
- Approved extensions

What are the costs*?

- Pay per instance
  - vCPU: $30.149 / vCPU / month
  - RAM: $5.11 / GB/ month
  - Storage: $0.17 / GB/ month
- High-availability virtually doubles your cost

Store Data – AlloyDB

What is it?

- “Disaggregated storage and compute”
- Cloud SQL Postgres but better
- Twice the performance of Postgres for same specs
- Good at both transactional and analytical workloads
- Vertex AI integration

What are the costs*?

- Pay per instance
  - vCPU: $48.2384 / vCPU / month
  - RAM: $8.176 / GB/ month
  - Storage: $0.299957 / GB/ month
- High-availability virtually doubles your cost
- AlloyDB Omni runs anywhere

Store Data – Spanner

What is it?

- Relational, graph, key-value, and search database on a global scale
- PostgreSQL interface
- Wide variety of data types
- Very expensive but enormously simplifies global-scale databases

What are the costs*?

- Pay for dedicated compute nodes
  - Compute capacity and storage
- “Data Boost” for on-demand workloads

Agenda

~~Run Code~~

- Compute Engine
- GKE
- Cloud Run
  - Services
  - Functions
  - Jobs

~~Store Data~~

- Cloud Storage
- Firestore
- Cloud SQL
- AlloyDB
- Cloud Spanner

Misc.

- Artifact Registry
- BigQuery
- Cloud Build

Misc. – Artifact Registry

What is it?

- Artifact storage (Docker, custom packages, etc.)
- Standard and remote modes
- Built-in vulnerability scanning (optional)
- Cleanup policies

What are the costs*?

- `Storage:  $0.10 / GB / month`
- `Scanning: $0.26 / image`

Misc. – BigQuery

What is it?

- Massive data analysis
- Queries are split across hundreds of compute nodes
- Serverless or dedicated compute
- GoogleSQL
  - SQLAlchemy plugin

What are the costs*?

- $0.023 / active GB / month
- $0.016 / inactive GB / month

Serverless:

- 2,000 slots available
- $6.25 / TB scanned

Dedicated compute:

- $0.04 / slot / hour

Misc. – BigQuery

Query data in...

- BigQuery
- BigTable
- Cloud Storage
- AWS S3
- Azure Blob Storage
- Google Drive
- Cloud SQL
- AlloyDB
- Spanner

Replicate data into BigQuery via Datastream

1. Address any known limitations
2. Set up a Datastream connector to your Cloud SQL database
3. Set up logical decoding in Postgres
4. Import schemas into BigQuery
5. Backfill data into BigQuery
6. Sit back and enjoy your CDC system

Misc. – Cloud Build

What is it?

- GitHub Actions but in GCP
  - Enable Cloud Build app in GitHub
  - Connect your repo to Cloud Build
  - Create a cloudbuild.yaml file
  - Create a trigger in Cloud Build
- Runs under a service account
- Access to private GCP services
- Build artifacts can be shared between jobs

What are the costs*?

- Standard workers
  - $0.003 / minute (1 vCPU, 4GB RAM)
  - $0.006 / minute (2 vCPUs, 8GB RAM)
    - 2,500 minutes free per month
  - $0.016 / minute (8 vCPUs, 8GB RAM)
  - $0.064 / minute (32 vCPUs, 32GB RAM)
- Private pools
  - $0.00317 / vCPU / minute
  - $0.00042 / GB RAM / minute

Agenda

~~Run Code~~

- Compute Engine
- GKE
- Cloud Run
  - Services
  - Functions
  - Jobs

~~Store Data~~

- Cloud Storage
- Firestore
- Cloud SQL
- AlloyDB
- Cloud Spanner

~~Misc.~~

- Artifact Registry
- BigQuery
- Cloud Build

Thank you!

Check out more GCP services here:
cloud.google.com/products