The Cloud Operator’s HUD

Architecture, Operations, and Observability for Google Cloud

```
gcloud compute environments describe --format=visual
```

Lens: Declarative
(RAD UI / Code)

Operator
(GCP Console / CLI)

Compute Operations

Provisioning, CI/CD, Traffic Splitting, Autoscaling

(VMs, Cloud Run, GKE)

Storage & Databases

Object Lifecycle, Backup Mechanics, Fleet Management

(GCS, Cloud SQL, Spanner)

Network Connectivity

Subnet Expansion, IP Reservation, Custom Routes

(VPCs, Peering, DNS)

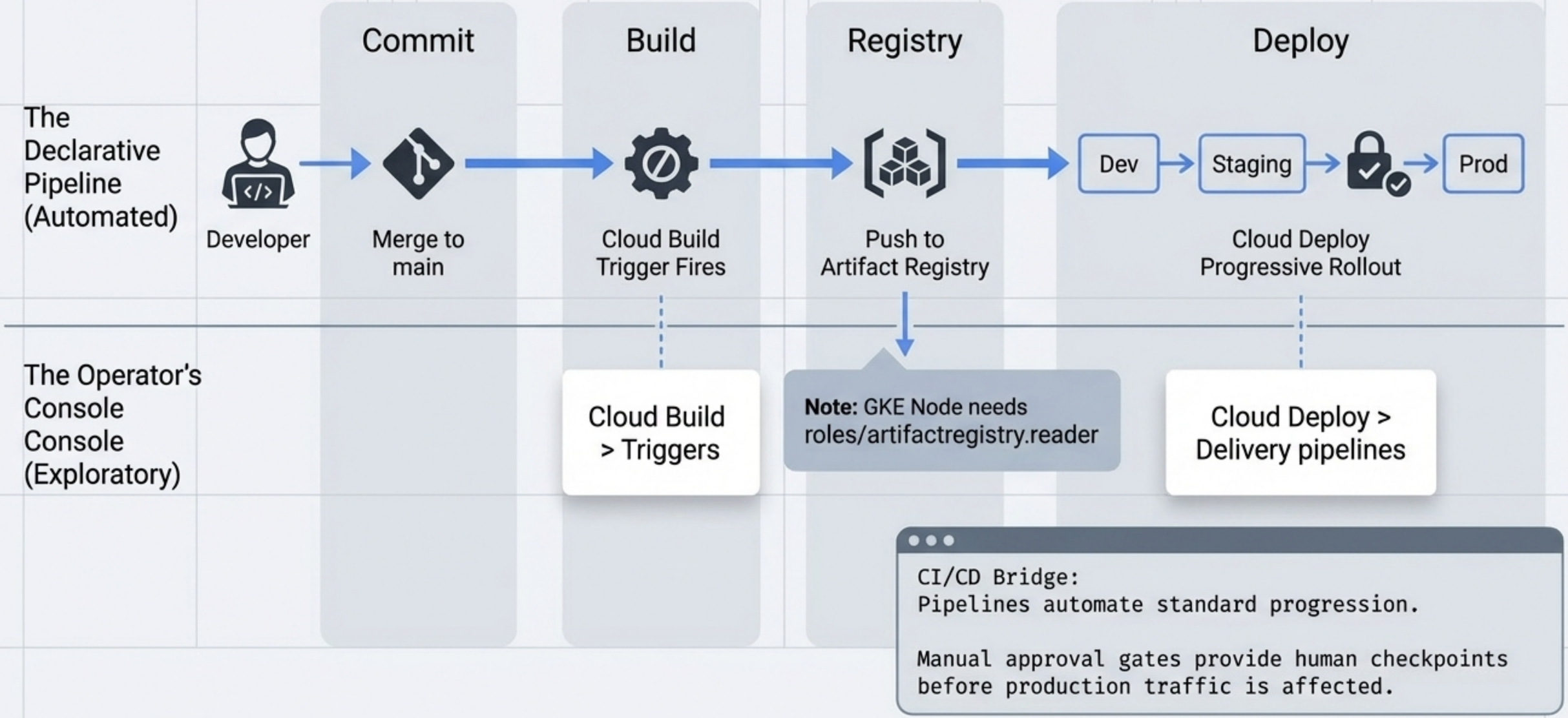
Observability

Log Routing, Threshold Alerting, Tracing, Profiling

(Ops Agent, Cloud Monitoring)

Mastering operations requires bridging the gap: defining infrastructure declaratively in code, and diagnosing it manually in the console.

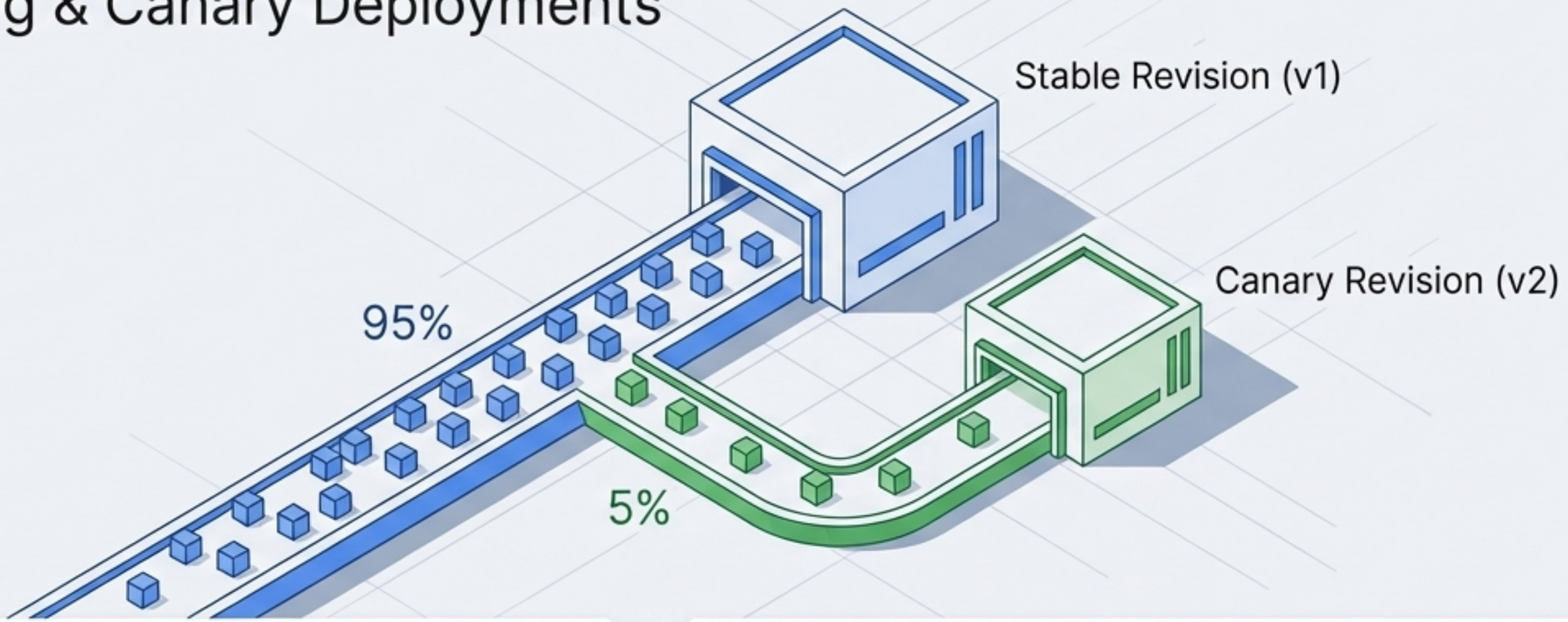
Automating the Delivery Pipeline




```
CI/CD Bridge:
Pipelines automate standard progression.

Manual approval gates provide human checkpoints
before production traffic is affected.
```

Traffic Splitting & Canary Deployments



Declarative View

```
Cloud Run:
  traffic_split = 5%
GKE:
  Multi-stage Cloud Deploy targets
```

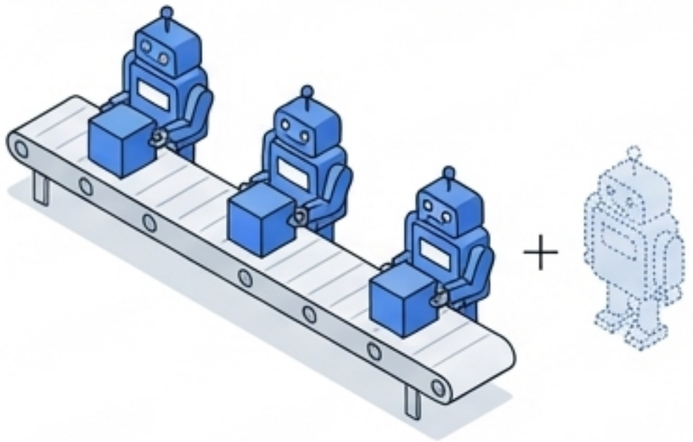
Operator View

```
Console Path:
  Cloud Run > Revisions > Traffic Column
```

The Operator's Insight: Zero-downtime rollbacks. If Cloud Monitoring detects an error spike on the 5% canary, an operator can instantly shift 100% of **traffic back to the stable revision** with a single variable change.

Kubernetes Scaling: Horizontal vs. Vertical

Horizontal Pod Autoscaler (HPA)



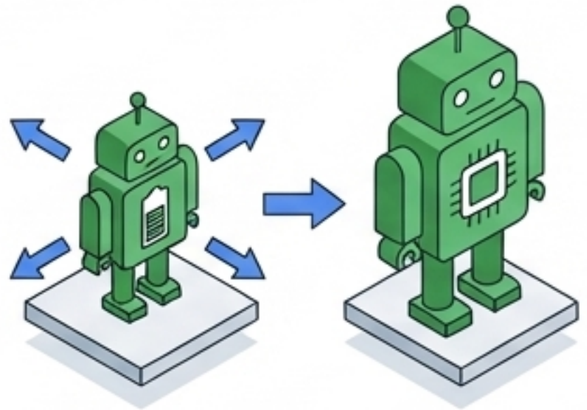
Mechanism: Adds more identical Pod replicas.

Trigger: CPU/Memory utilization targets.

```
kubectl scale deployment <name> --replicas=3
```

```
min_instance_count
max_instance_count
```

Vertical Pod Autoscaler (VPA)



Mechanism: Increases CPU/Memory requests of existing Pods.

GKE Autopilot Note: Bills per pod resource request. Right-sizing directly controls cost.

```
Explicit container resource requests are strictly required.
```

Reliability Guardrail: Pod Disruption Budgets (PDBs) ensure a minimum number of replicas remain available during voluntary node disruptions.

Compute Access & Scaling Diagnostic Matrix

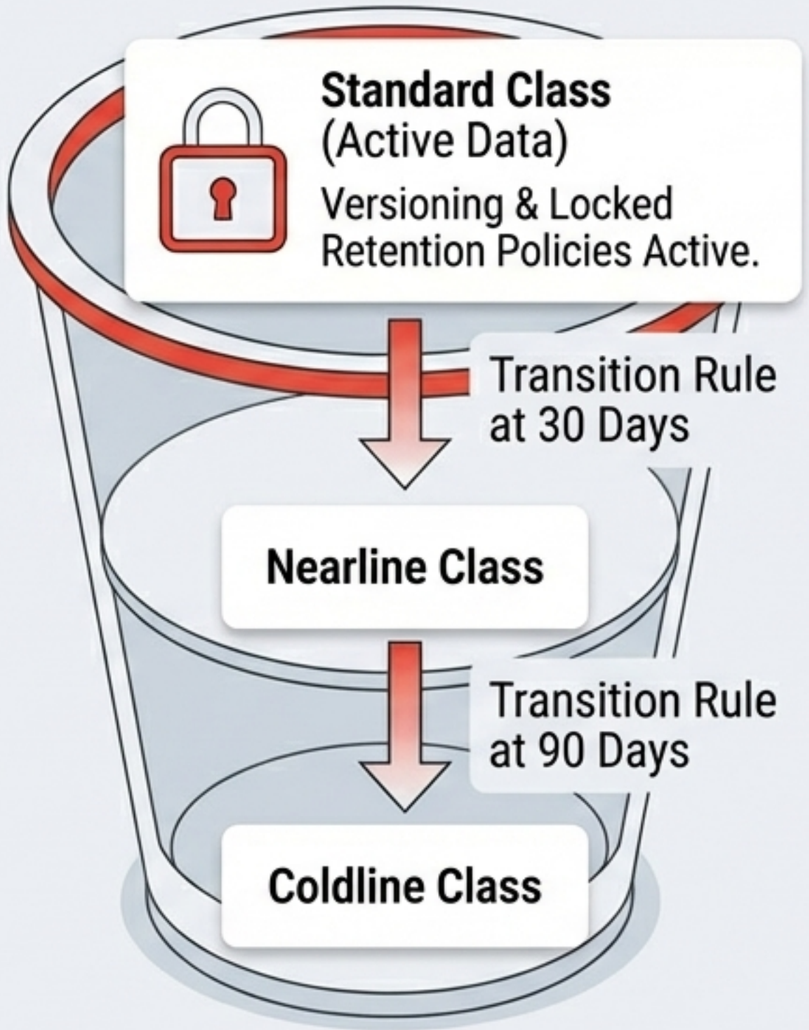
Compute Engine (VMs & MIGs)	Cloud Run (Serverless)	GKE (Autopilot/Standard)
Scaling Trigger: CPU Utilization via MIG policies	**Scaling Trigger:** Concurrent requests.	**Scaling Trigger:** HPA/VPA via CPU/Memory resource requests.
Operator Access: OS Login with IAP TCP Forwarding (No public IP needed)	**Cold Start Note:** Setting min instances > 0 prevents cold starts.	**Operator Access:** Managed via Control Plane API.
`gcloud compute ssh --tunnel-through-iap`	**Operator Access:** Fully managed; invoked via API/HTTPS.	`kubectl get pods -A`
State Management: Snapshots = Point-in-time disk backup. Images = Portable OS for new VMs.	**State Management:** Strictly stateless ephemeral containers.	**State Management:** StatefulSets provide stable network identities (<name>-0) and dedicated PVCs.

Cloud Storage: Lifecycle & Protection

The Security Shield

Uniform Bucket-Level Access: Recommended. IAM-only permissions.

Fine-Grained Access: Legacy object-level ACLs.



The Transfer Engine

Storage Transfer Service

Use Case: Large-scale, cross-region syncing (e.g., Disaster Recovery buckets).

Advantage: Built-in checksums and automatic retry logic. Supersedes manual `gsutil cp` commands.

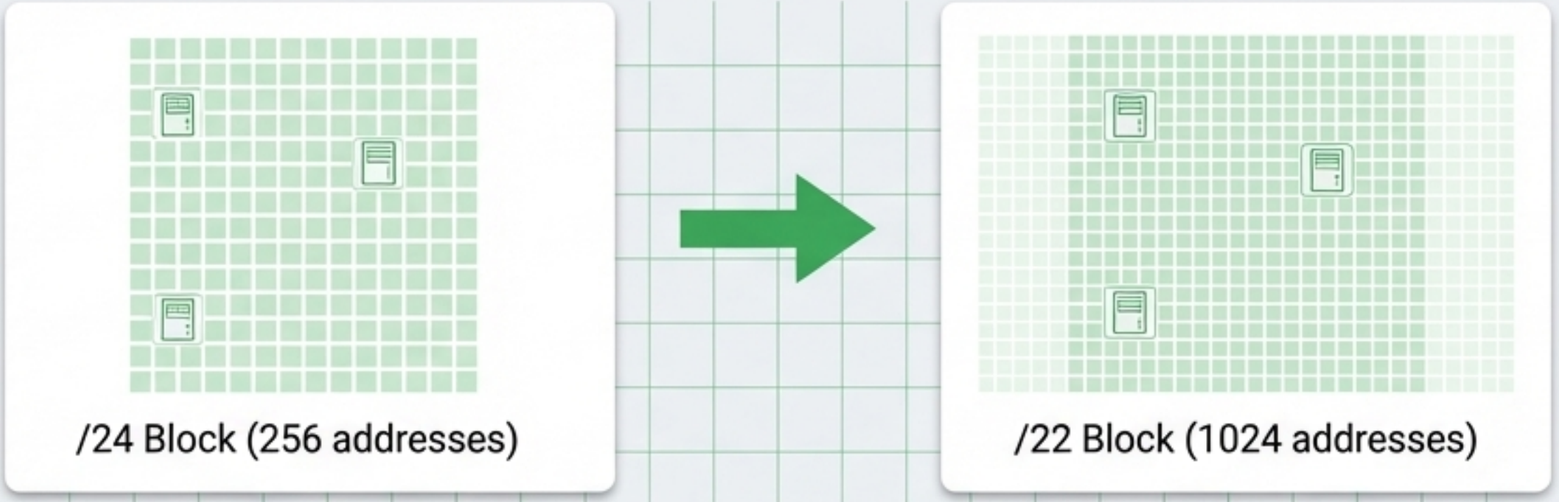
GCP Database Fleet Diagnostic Matrix

	Cloud SQL	AlloyDB	Spanner	Firestore	Bigtable
Backup	Point-in-Time Recovery (PITR) up to 35 days.	Continuous with 14-day PITR.	Online backups without downtime.	Scheduled Import/Export to GCS.	In-service snapshots.
Query	`gcloud sql connect` (`Cloud Shell`)	`PostgreSQL` client in Cloud Shell	`Spanner Studio` in Console	Firestore Data Viewer	`cbt read <table-name>` or Bigtable Studio

Database Center

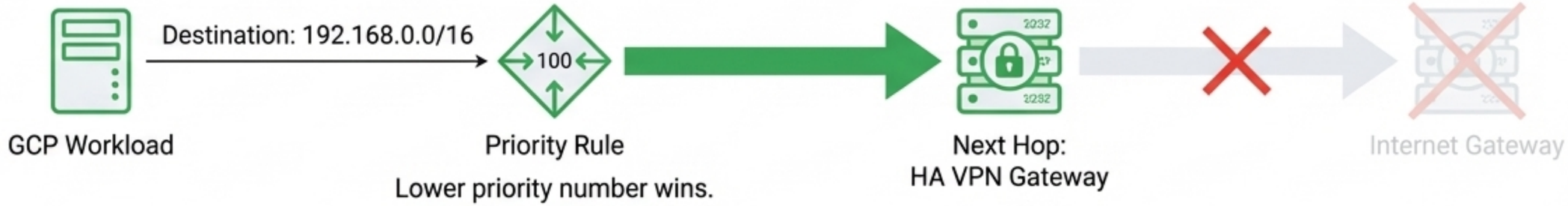
A unified pane for fleet management, surfacing Recommender and Security Command Center insights across all database products.

The Subnet Expansion Model



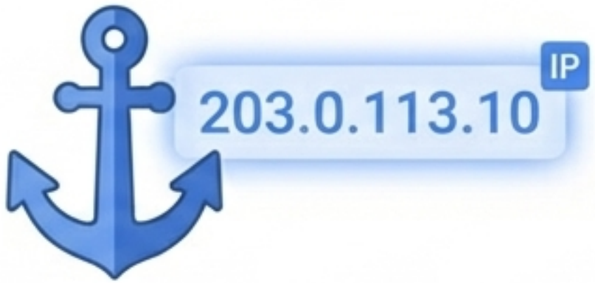
Rule: CIDR expansions are permanent and non-disruptive. Existing IPs do not change. You cannot shrink a subnet once expanded.

Custom Static Routes



Network Connectivity Rules

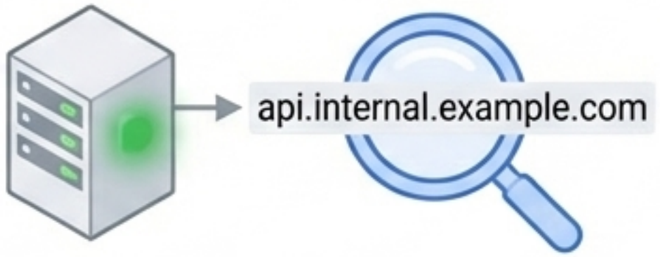
IP Reservation



External Global IPs: Strictly required for Global External Application Load Balancers.

Internal Static IPs: Provide stable, unchanging references for internal resources like databases.

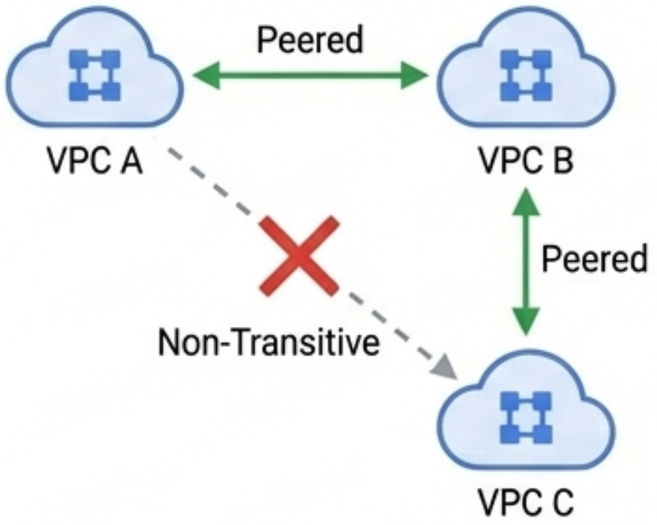
Cloud DNS (Private Zones)



Keeps resolution entirely inside the VPC.

Allows internal microservices to reference each other by stable hostname rather than fragile IP addresses.

VPC Peering Rules



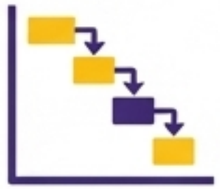
Rule Explicit: VPC Peering is strictly non-transitive. A cannot reach C through B.

The Observability Stack Matrix

Cloud Monitoring

Signal: `Metrics / Alerts`

Use Case: Uptime checks, custom log-based metrics, threshold alerts (e.g., p95 latency, `CrashLoopBackOff`).



Cloud Trace

Signal: `Latency`

Use Case: End-to-end request latency tracking across microservices via `OpenTelemetry` spans.

Cloud Profiler

Signal: `Code-level CPU / Memory`

Use Case:
Identifies function-level bottlenecks in production code with <1% overhead.

Query Insights

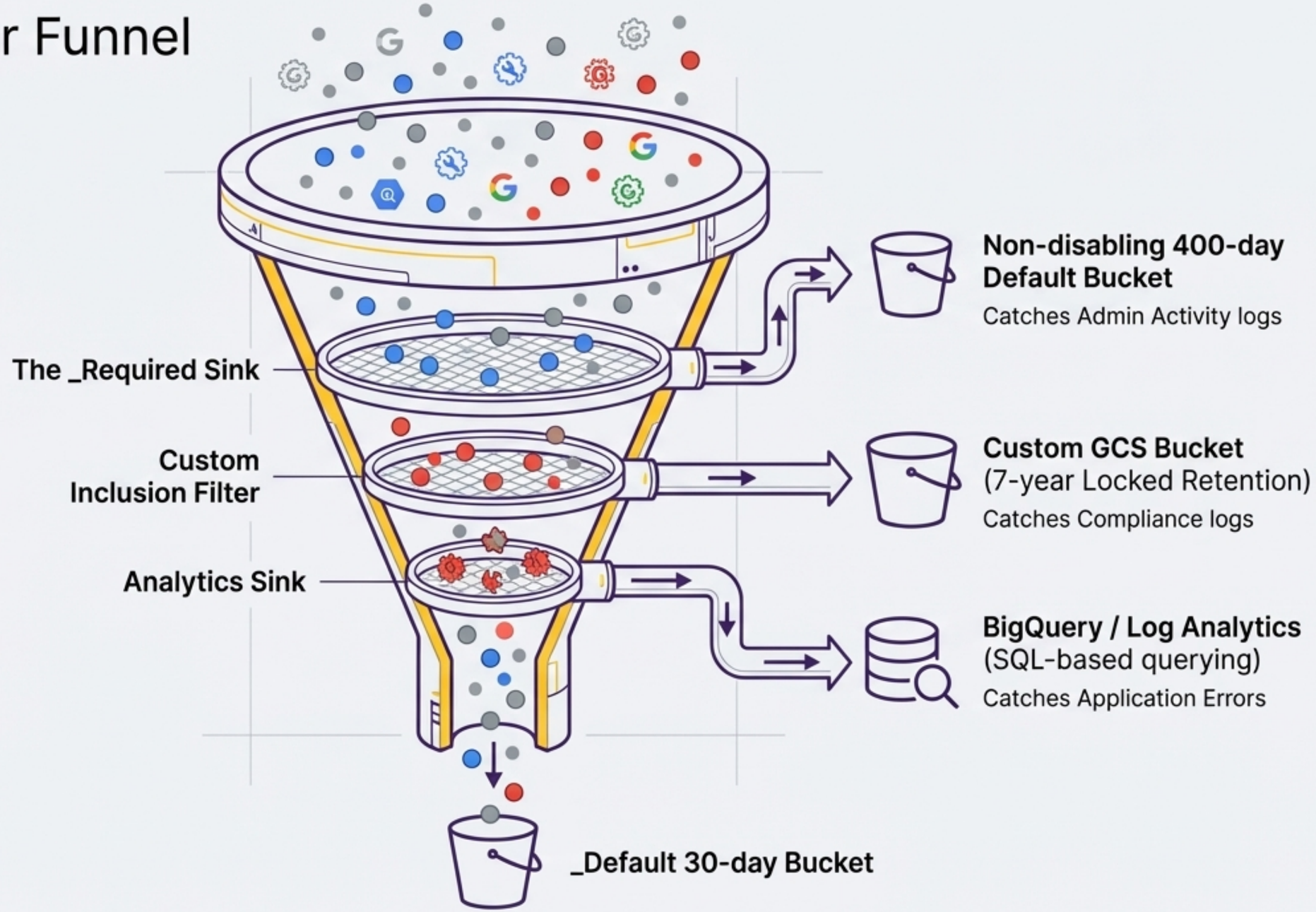
Signal: `Database Load`

Use Case:
Pinpoints high-load `Cloud SQL` queries and displays `EXPLAIN ANALYZE` execution plans.

The Telemetry Collectors: Compute Engine VMs utilize the unified Ops Agent, while GKE Autopilot utilizes Managed Service for Prometheus seamlessly.

The Log Router Funnel

Note: Data Access audit logs are OFF by default and must be manually enabled. Logs can also be exported to on-premises SIEM tools via Pub/Sub topics.



Proactive Diagnostics & Insights

Recommendation

Active Assist (Recommender)

Idle VMs: Identify low-usage compute for deletion.

Overprovisioned VMs: Downsize machine types based on historical CPU data.

Unused IPs: Release orphaned static IP reservations to save costs.

IAM Policy Insights: Automatically flag and remove roles unused in the last 90 days.

Recommendation

Personalized Service Health

Compute Engine	Cloud Storage
Cloud Storage	BigQuery
GKE	Cloud Storage
Cloud Resource	BigQuery
BigQuery	

Filters the massive global GCP status page.

Displays only the specific 8-12 services actively deployed in your project.

Reduces noise during incident response.

Recommendation

Gemini Cloud Assist

Natural language query:

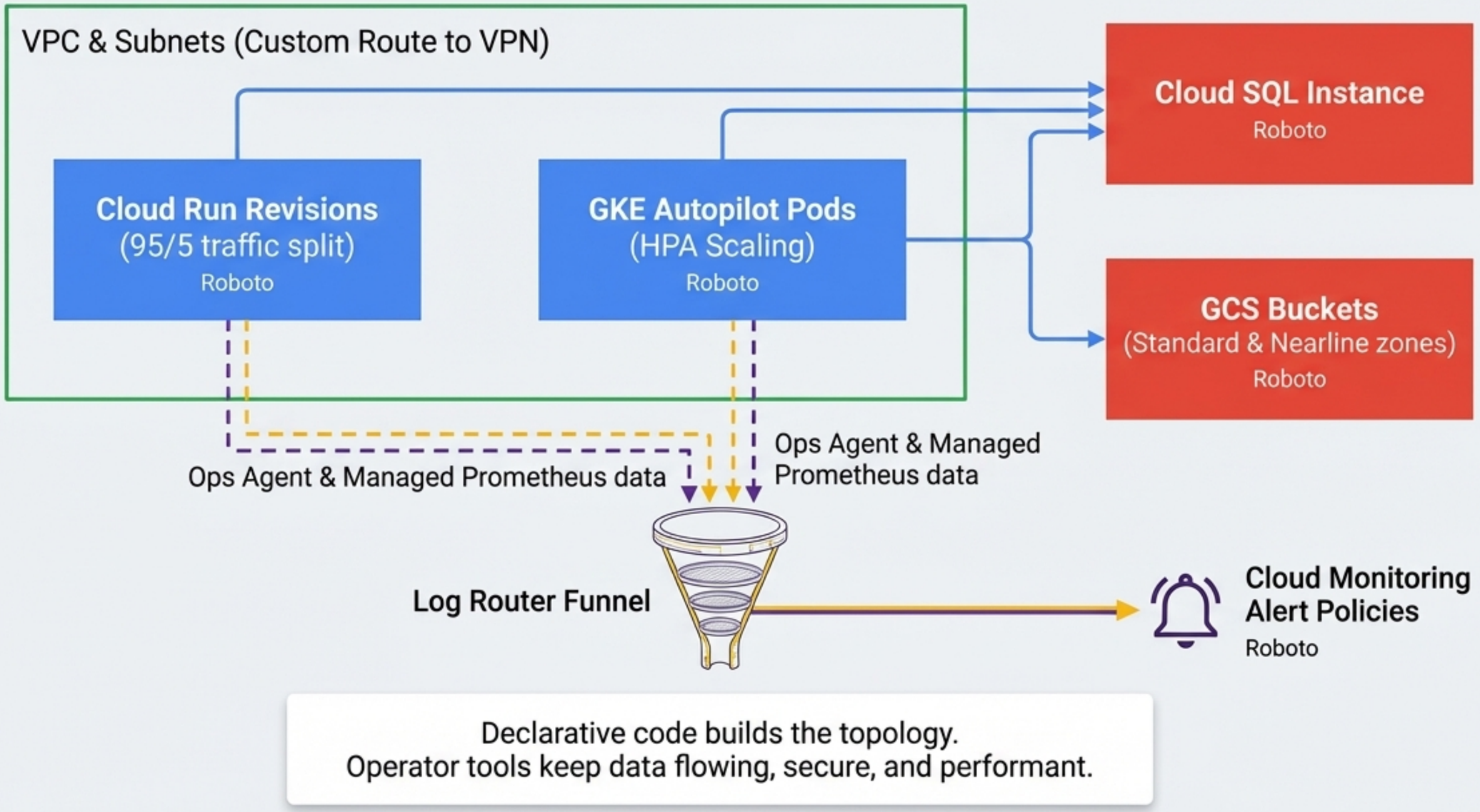
Ask a question...

Natural language querying for monitoring data.

Example prompt: “Which Cloud Run services have the highest p95 latency this week?”

Assists in writing complex MQL (Monitoring Query Language) expressions.

The Unified Operations Model



Declarative code builds the topology.
Operator tools keep data flowing, secure, and performant.

The Operator’s Cheat Sheet

Access & Discovery

Shell & Inventory	`gcloud compute ssh <vm> --tunnel-through-iap` (Secure VM shell)	`kubectl get pods -A` `kubectl get services` (GKE inventory)	`gcloud sql connect <instance>` (Cloud SQL shell)

Cost & Analysis

Dry Runs & Queries	`bq query --dry_run` (Estimate BigQuery bytes/cost before running)	`bq ls -j` (List BigQuery jobs & history)	`cbt read <table-name>` (Bigtable query)

Operations & Scale

Modifying State	`gcloud compute networks subnets expand-ip-range <name> --prefix-length=22`	`kubectl scale deployment <name> --replicas=X`

ACE Section 3 Quick Reference