GCP

Professional Cloud Architect

Agenda

1. **Introduction**
2. **Business Logic Design**
3. **Data Layer Design**
4. **Network Layer Design**
5. **Design for security**

Introduction

What is a Cloud Architect?

- **A cloud Architect enables organizations to leverage...Cloud technologies.**

  Through an understanding of cloud architecture and ...technology, this individual designs, develops, and manages robust, secure, scalable, highly available, and dynamic solutions to drive business objectives.

- **The Cloud architect should be proficient in all aspects of enterprise cloud strategy, solution design, and architectural best practices.**
- **The cloud Architect should also be experienced in software development methodologies and approaches including multi-tiered distributed applications which span multi-cloud or hybrid environments.**

Key abilities

The Cloud Architect [should have the] ability to:

- **Design and plan a cloud solution architecture**
- **Manage and provision the cloud solution infrastructure**
- **Design for security and compliance**
- **Analyze and optimize technical and business processes**
- **Manage implementations of cloud architecture**
- **Ensure solution and operations reliability**

What are “Systems”?

- The “Things” we architect and build with cloud tools
- Always about data flows:
  - Moving information
  - Remembering information
  - Processing information
- Driven by a purpose
  - To achieve something
  - To solve a problem

Not just "Architecting Great Systems"

Not just great *systems*, when they're in use...

Nor just great *architectures*, when they're complete...

But also great *architecting*, while it happens...

And great *building* of those architectures.

The *process* matters.

A lot.

Positionally

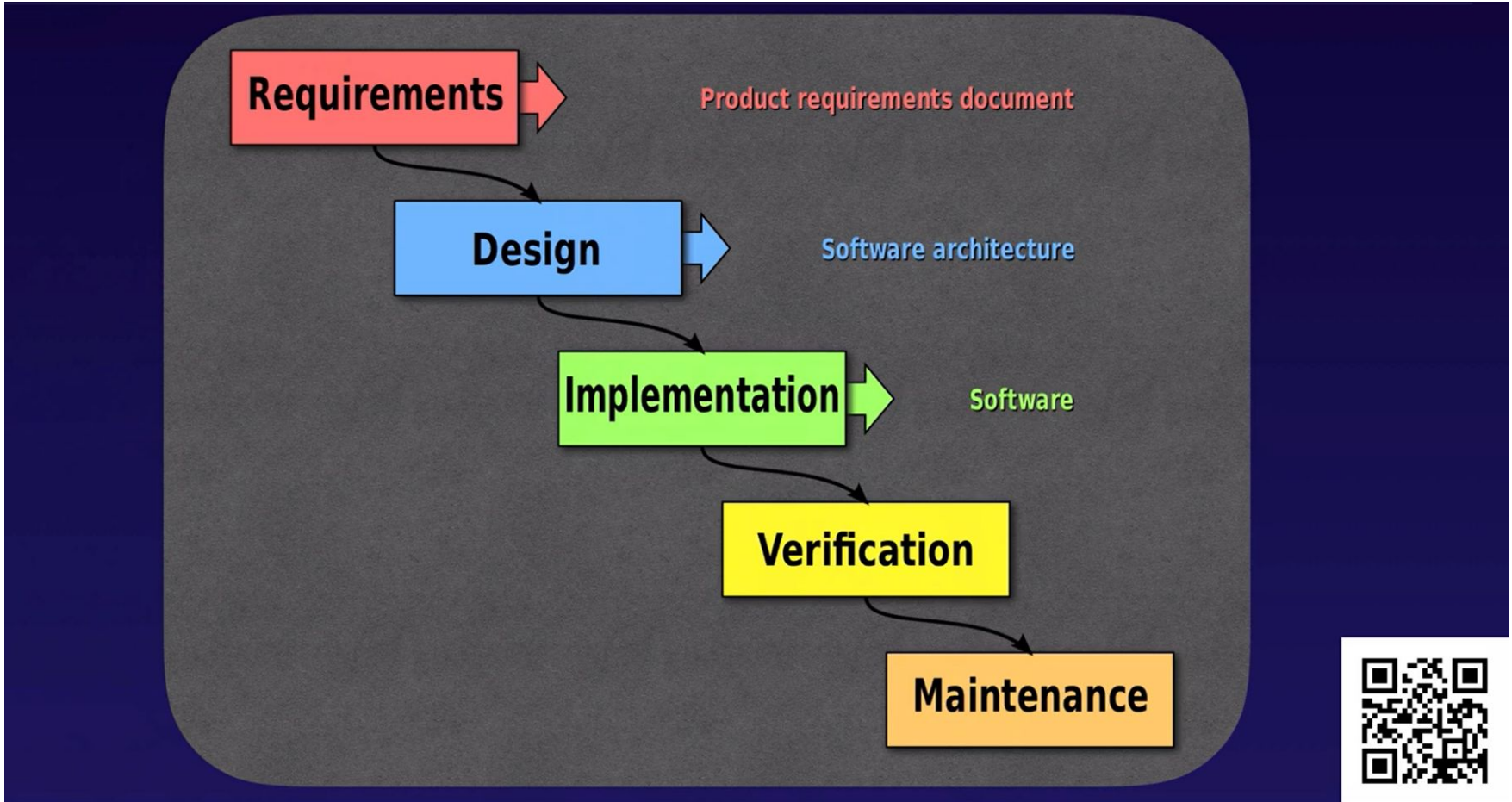
Cloud Architect

is a

Leadership

Position

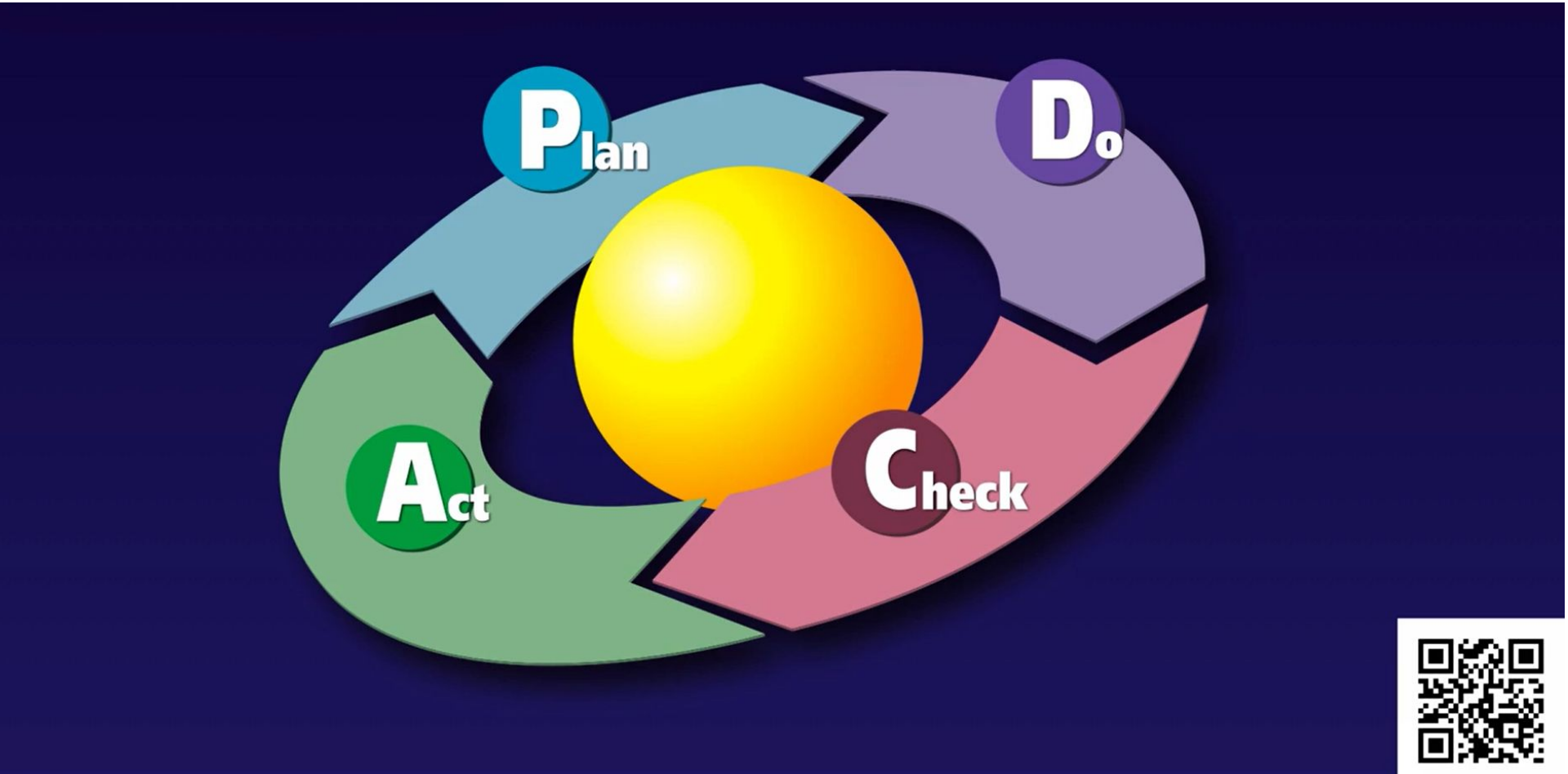
Waterfall


Requirements
Product requirements document
Design
Software architecture
Implementation
Software
Verification
Maintenance

Agile


Manifesto for Agile Software Development
We are uncovering better ways of developing
software by doing it and helping others do it.
Through this work we have come to value:
Individuals and interactions over processes and tools
Working software over comprehensive documentation
Customer collaboration over contract negotiation
Responding to change over following a plan
That is, while there is value in the items on
the right, we value the items on the left more.

Iteration: Deming Cycle/PDCA



Strategic Stuff

How to Plan:

1. Understand your environment
2. Determine where you are now
3. Figure out where you want to go
4. Decide how you'll get there

Doing the Architecting Thing

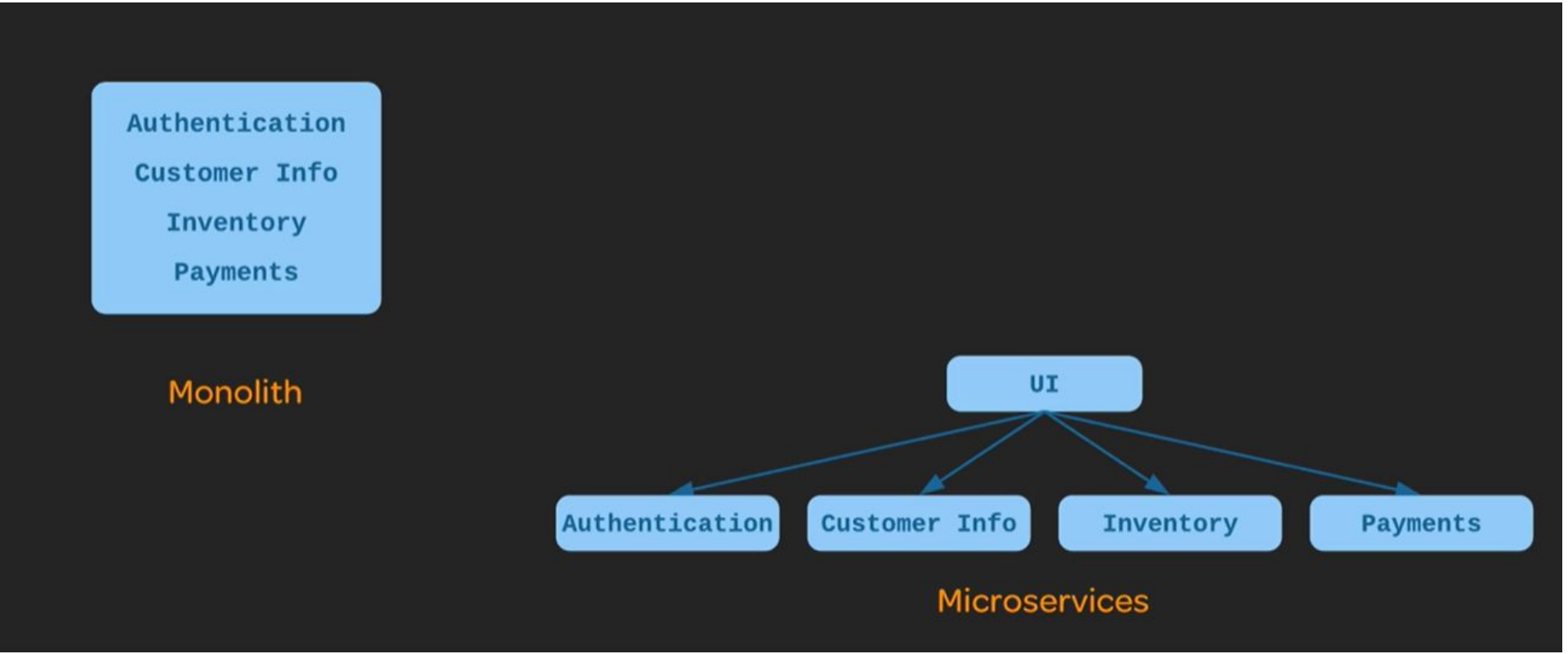
Foundation	How does this stuff work?
Environment	What affects us?
Current	Where are we?
Target	Where do we want to be?
Plan	How will we get from here to there?
Execute	Let's go!
Evaluate	Are we getting there?
Adjust	What should we change?

Business Logic Design

What are Microservices?

- Microservices: A microservice architecture breaks an application up into a collection of small, loosely-coupled services
- Traditionally, apps used a monolithic architecture. In a monolithic architecture, all features and services are part of one large application
- Microservices are small: each microservice implements only a small piece of an application's overall functionality
- Microservices are loosely coupled: Different microservices interact with each other using stable and well-defined APIs. This means that they are independent of one another

Microservices vs Monolith



What do microservices look like?

- There are many different ways to structure and organize a microservice architecture

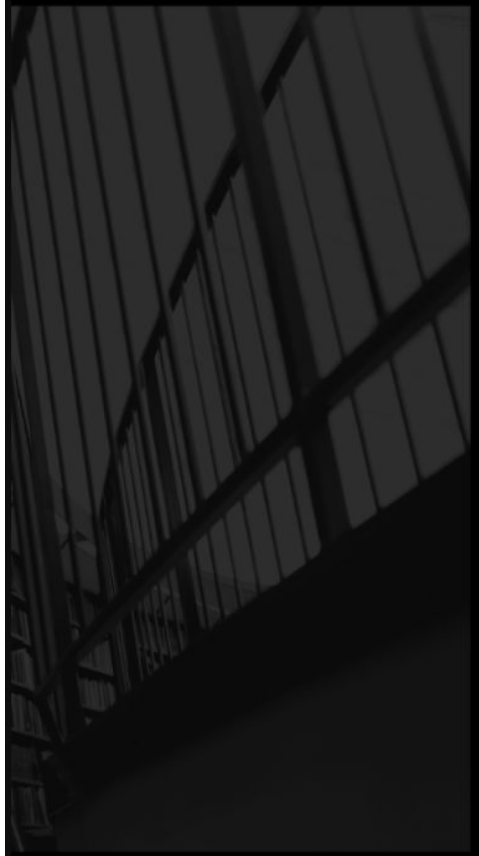
- For example, a pet shop application might have:
  - A pet inventory service
  - A customer details service
  - An authentication service
  - A pet adoption request service
  - A payment processing service

- Each of these is its own codebase and a separate running process (or processes). They can all be built, deployed, and scaled separately

Why use Microservices?

- Modularity – Microservices encourage modularity. In monolithic apps, individual pieces become tightly coupled, and complexity grows. Eventually, it's very hard to change anything without breaking something
- Technological flexibility – You don't need to use the same languages and technologies for every part of the app. You can use the best tool for each job
- Optimized scalability – You can scale individual parts of the app based upon resource usage and load. With a monolith, you have to scale up the entire application, even if only one aspect of the service actually needs to be scaled
- Microservices aren't always the best choice. For smaller, simpler apps a monolith might be easier to manage

The Twelve-Factor App:

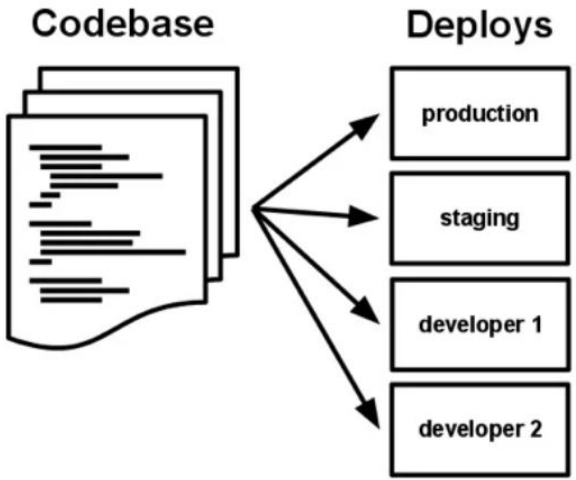


1. Codebase
2. Dependencies
3. Config
4. Backing Services
5. Build, Run, Release
6. Stateless Processes
7. Port Binding
8. Concurrency
9. Disposability
10. Dev-Prod Parity
11. Logs
12. Admin Processes

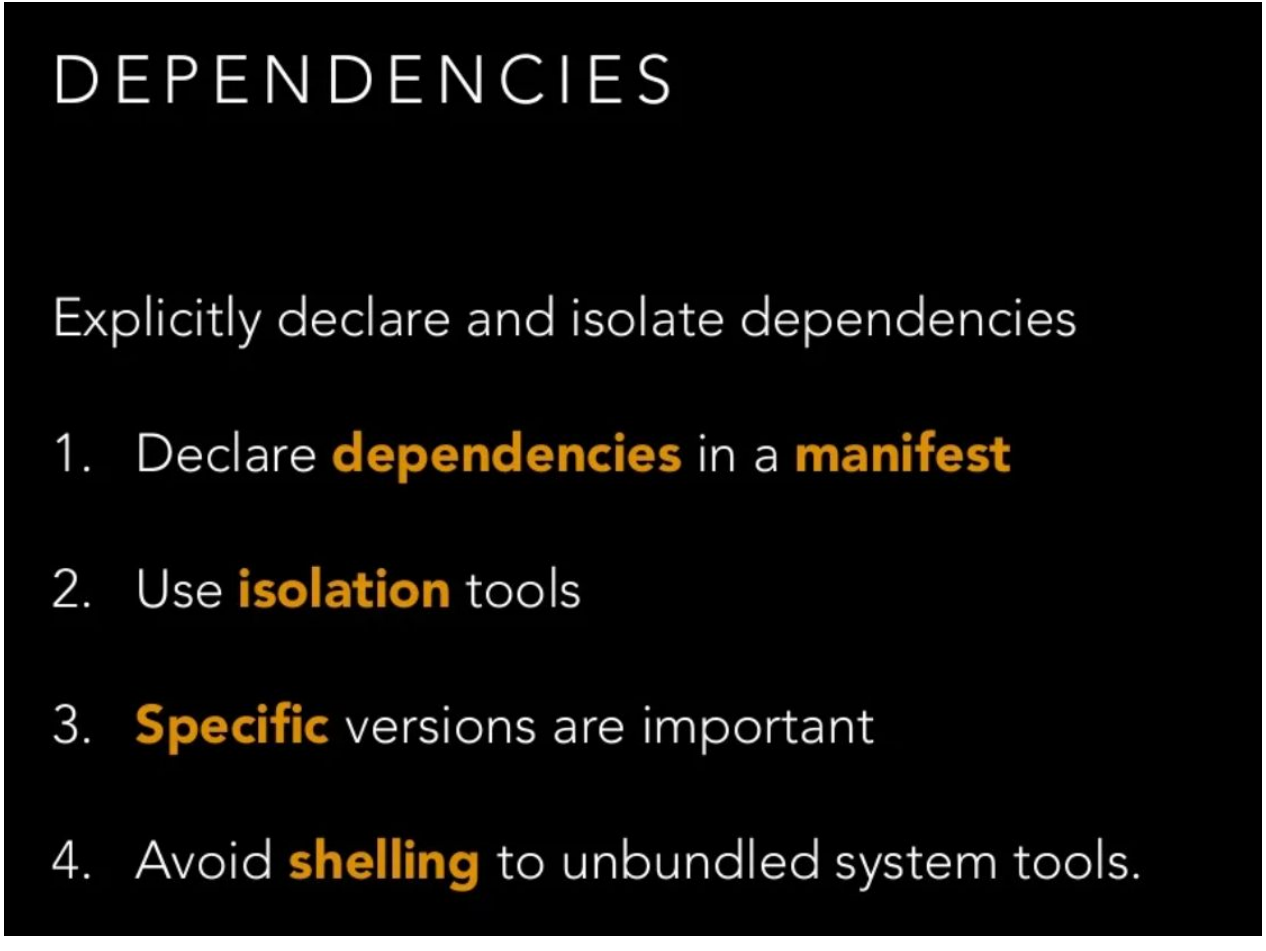
The twelve factors - Codebase

CODEBASE

- codebase = repo
- **one** repo - **many** deploys
- app **!=** many repos
- many repos = distributed system



The twelve factors - Dependencies



The twelve factors - Configuration

- Config is the **specific** information required to host a deployment of your codebase
- Does not include things like routes etc.
- Mainly database **credentials**, **paths**, **resource urls** etc.

The twelve factors - Configuration

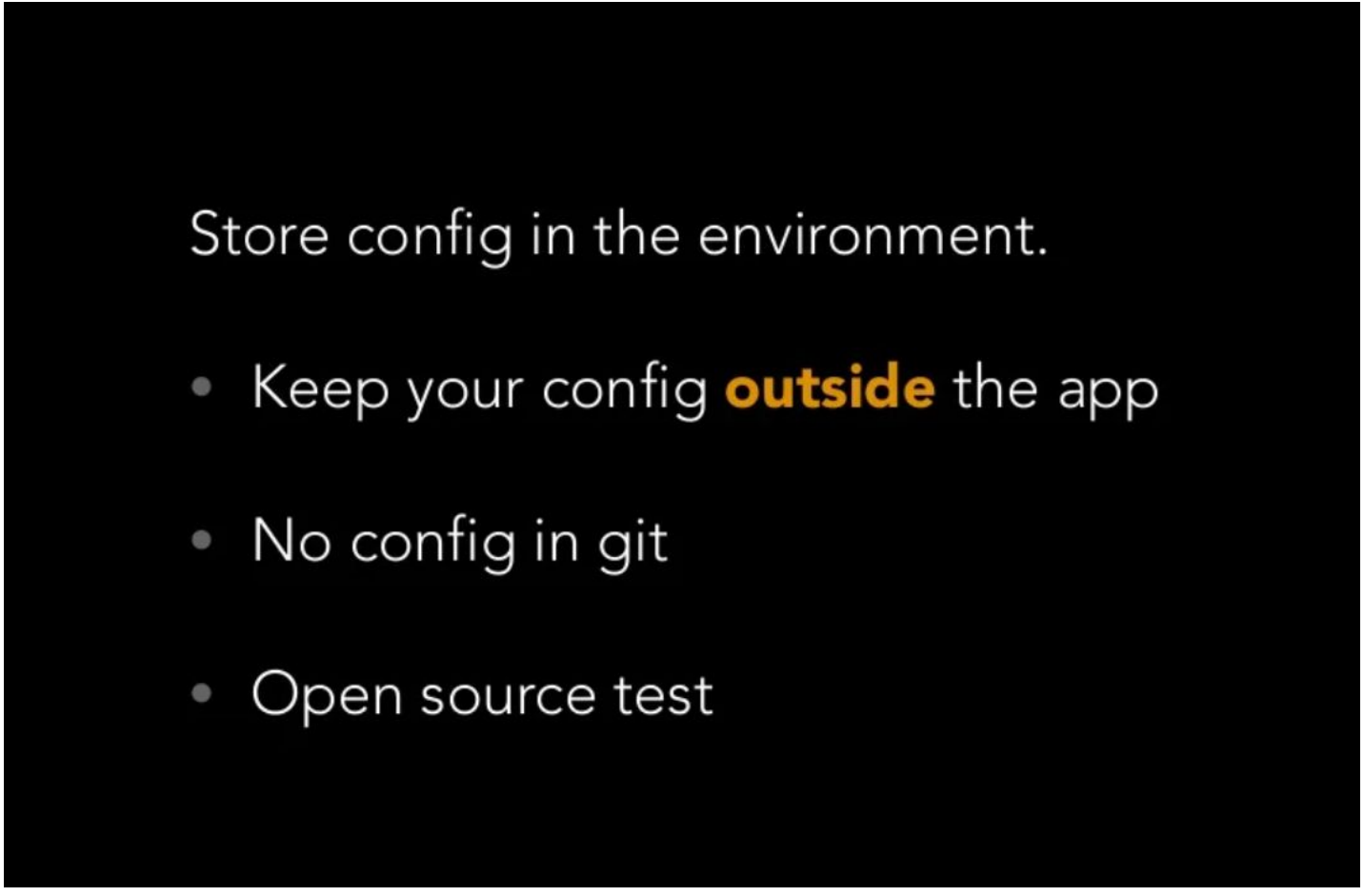
Examples:

Dependency Manifest = **Gemfile**

Isolation tools = **bundle exec**

```
gem "redis-rails", "~> 3.2.3"
```

The twelve factors - Configuration



The twelve factors - Configuration

Use **environment vars**

Can you make your repo open source today?

```
ENV['DATABASE_URL']
```

```
production:
  adapter: mysql2
  database: <%= ENV['DB_ENV_MYSQL_DATABASE'] %>
  username: <%= ENV['DB_ENV_MYSQL_USER'] %>
  password: <%= ENV['DB_ENV_MYSQL_PASS'] %>
  port: <%= URI(ENV['DB_PORT']).port %>
  host: <%= URI(ENV['DB_PORT']).host %>
```

The twelve factors - Backing services

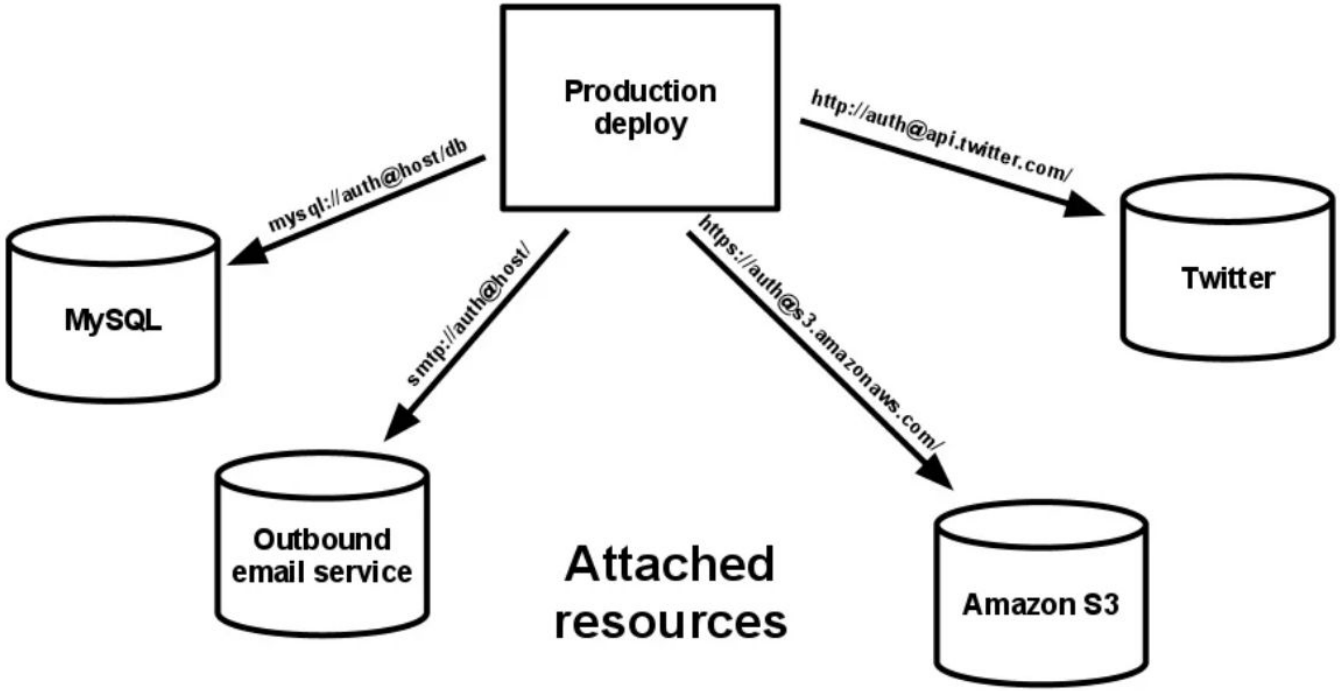
Treat backing services as attached resources

What's a backing service?

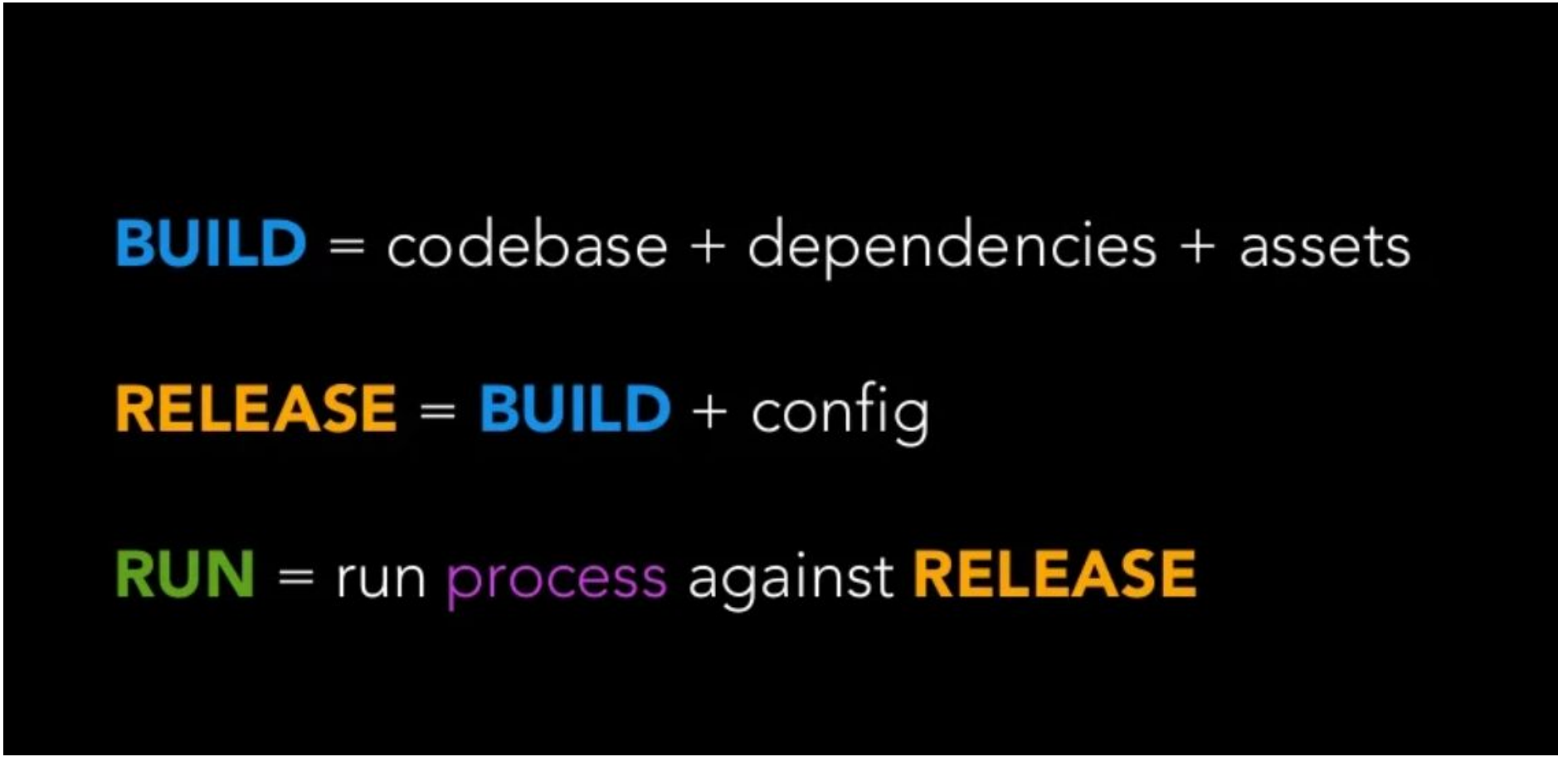
- **Datastore**
- **SMTP**
- **Caching systems**
- **Amazon S3**

Make no distinction between local and third party services

The twelve factors - Backing services



The twelve factors - Build, release, run



The twelve factors - Build, release, run

- Strict **separation** between stages
- **Cannot change** code at **runtime**
- **Rollback** = just use the last release instead.
- Release has unique release **ID**

The twelve factors - Build, release, run

- Does the **running** of the **release**
- Is **stateless**
- **Shares nothing** with other processes
- Uses **single transaction** only caches
- Session **db storage** …over sticky sessions
- Asset **pre-compilation** …over runtime calculation

The twelve factors - Processes

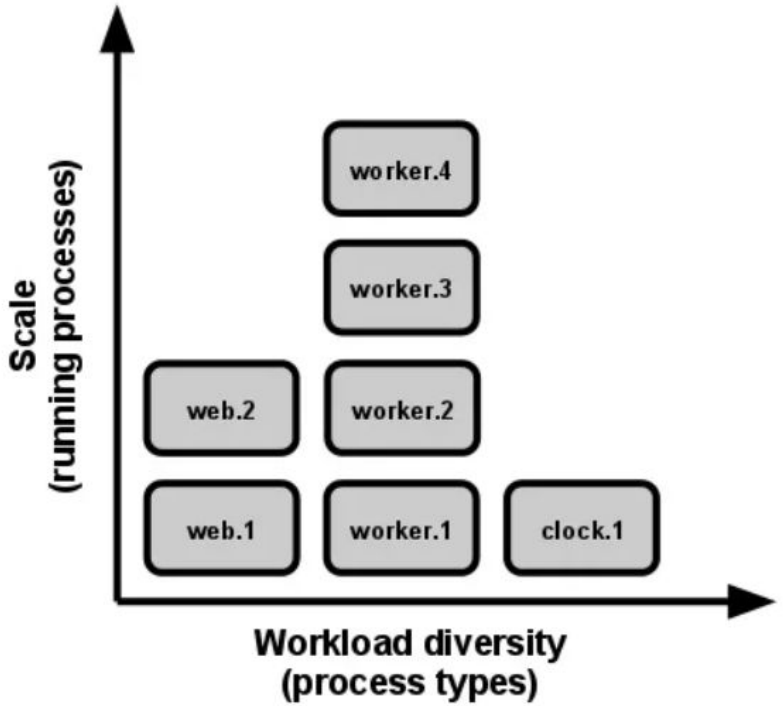
The twelve factors - Port binding

The twelve factors - Concurrency

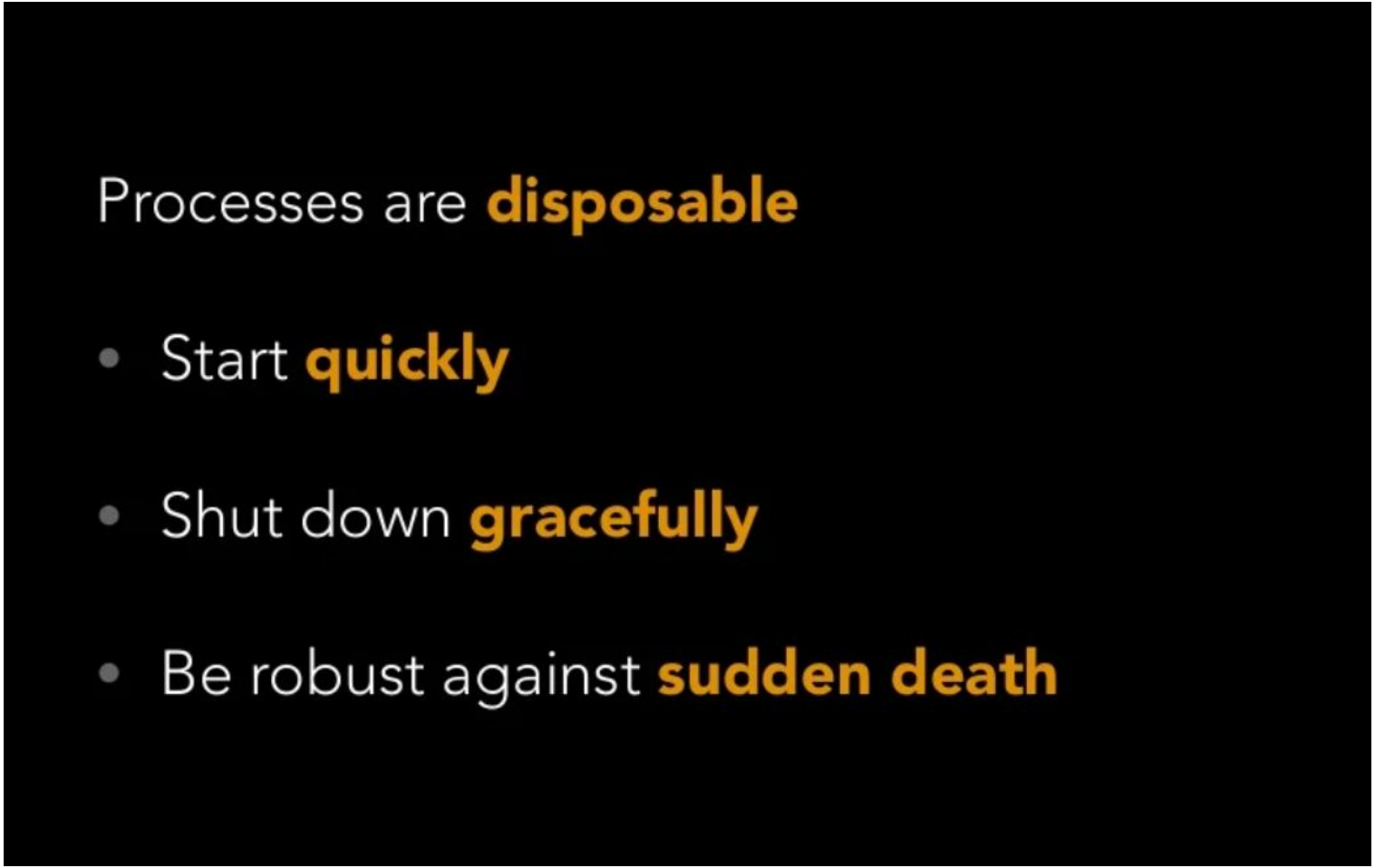
- **Scale out** via the **process model**
- Processes are **first class** citizens.
- Assign work to a process **type** (web, db, worker etc.)
- Can then handle **own multiplexing**
- Processes **don't daemonize** or write PID files.
- Instead, use system **process manager**

The twelve factors - Concurrency

Scale out by running multiple processes of different types



The twelve factors - Disposability

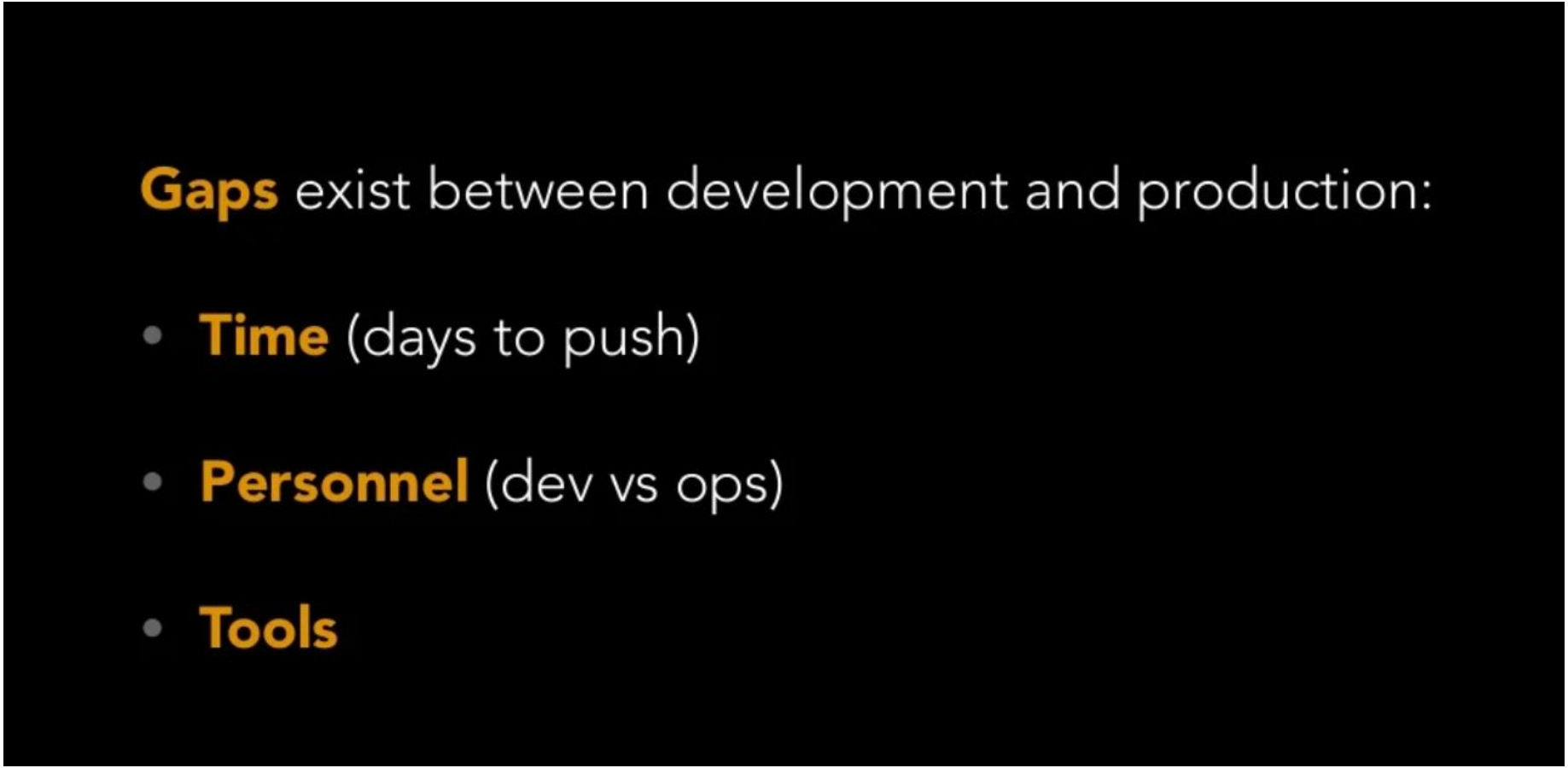


The twelve factors - Disposability

Example: Worker Process

- Return the current job to the job queue
- All jobs are **reentrant**
- Or at least **idempotent**

The twelve factors - Environment parity



The twelve factors - Environment parity

Need to **shorten the gaps!**

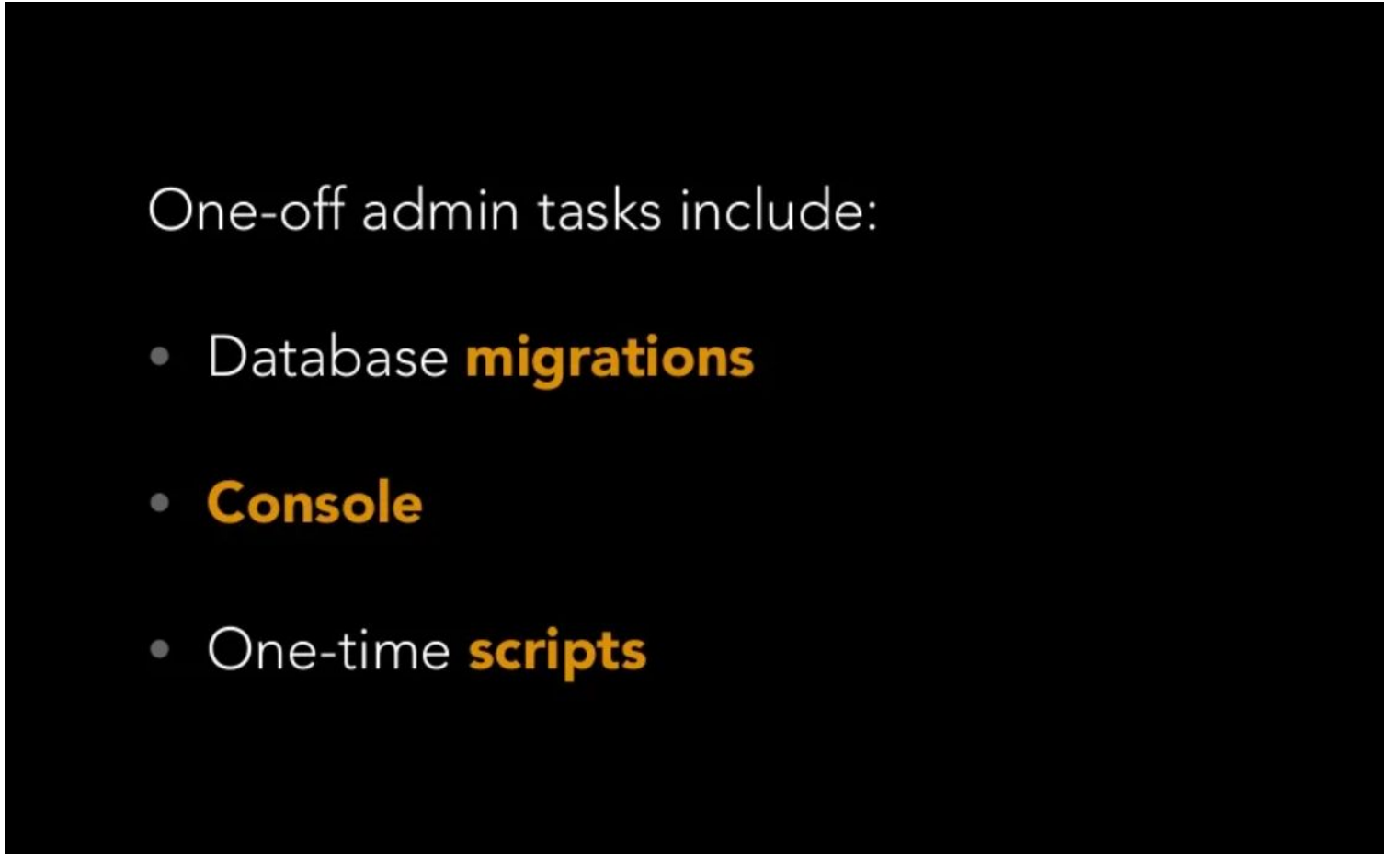
- CI & deploy **ASAP** after writing code
- Get Developers **involved** in Operations
- Environments should be as **similar** as possible:
  - Eg. **Resist the urge** to use **different** backing services between development and production

The twelve factors - Logs

Within your app treat logs as **event streams**

- **Don't** route or store **logs in files**
- **Stream** to stout instead and be captured and handled by the **environment**

The twelve factors - Admin processes

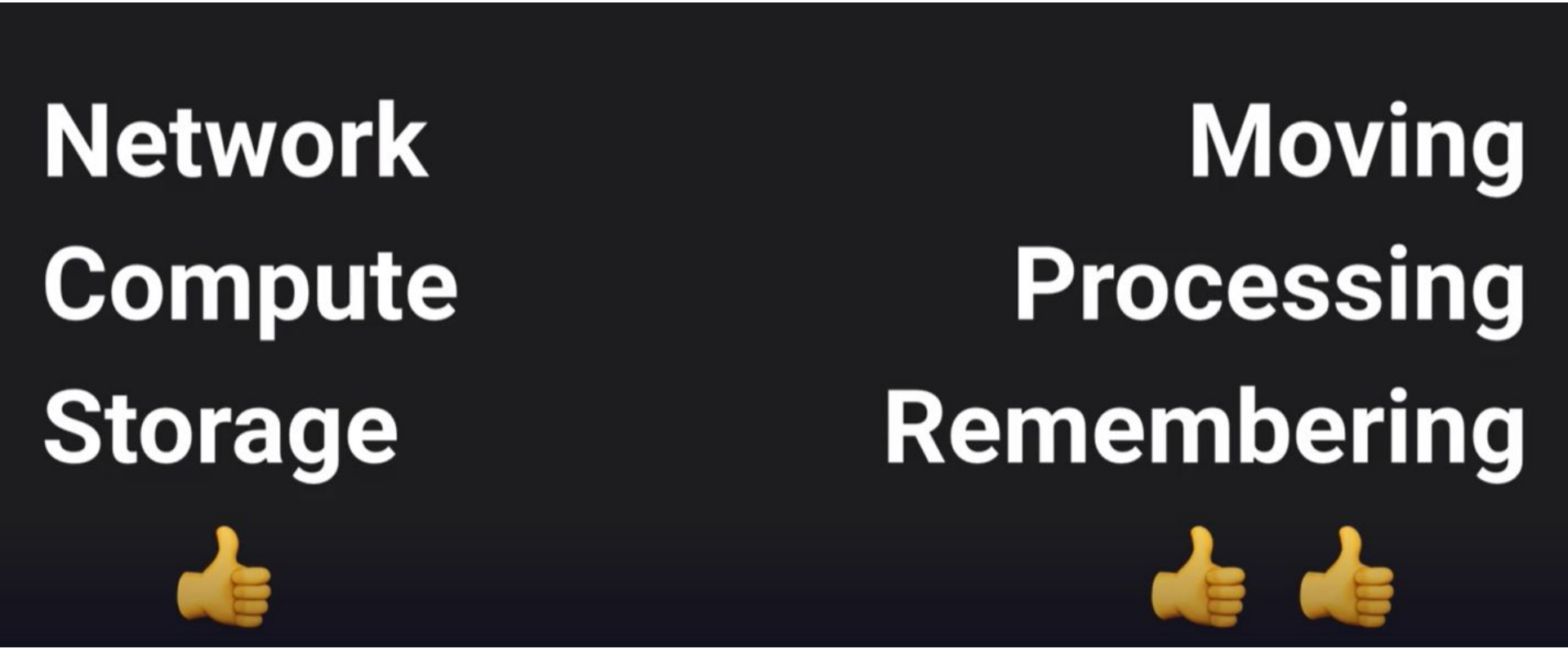


The twelve factors - Admin processes

Mapping computational needs for Google Cloud Platform processing services

Data Layer Design

Data flow foundations



Data Flows - Processing:

Data Flows - Processing: Compute Engine (GCE)

Zonal Regional Multi-Regional Global

- Fast-booting Virtual Machines (VMs) you can rent, on demand
- Infrastructure as a Service (IaaS)
- Pick set machine type—standard, highmem, highcpu—or custom CPU/RAM
- Pay by the second (60 second min.) for CPUs, RAM
- Automatically cheaper if you keep running it (“sustained use discount”)
- Even cheaper for “preemptible” or long-term use commitment in a region
- Can add GPUs and paid OSes for extra cost
- Live Migration: Google seamlessly moves instance across hosts, as needed

Data Flows - Processing: Kubernetes Engine (GKE)

Zonal Regional Multi-Regional Global

- Managed Kubernetes cluster for running Docker containers (with autoscaling)
- Used to be called “Google Container Engine” (but still GKE) until Nov, 2017
- Kubernetes DNS on by default for service discovery
- No IAM integration (unlike AWS’s ECS)
- Integrates with Persistent Disk for storage
- Pay for underlying GCE instances
  - Production cluster should have 3+ nodes
- No GKE management fee, no matter how many nodes in cluster, as of Nov 2017

Data Flows - Processing: App Engine (GAE)

App Engine

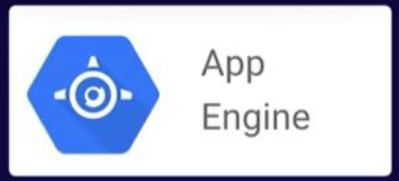
Zonal Regional Multi-Regional Global

- Platform as a Service (PaaS) that takes your code and runs it

Elastic Beanstalk

heroku

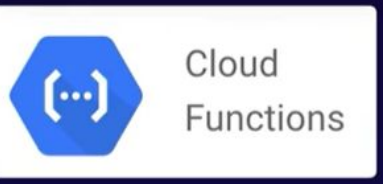
Data Flows - Processing: App Engine (GAE)



Zonal Regional Multi-Regional Global

- Platform as a Service (PaaS) that takes your code and runs it
- Much more than just compute — Integrates storage, queues, NoSQL, ...
- Flex mode ("App Engine Flex") can run any container & access VPC
- Auto-scales based on load
  - Standard (non-Flex) mode can turn off last instance when no traffic
- Effectively pay for underlying GCE instances and other services

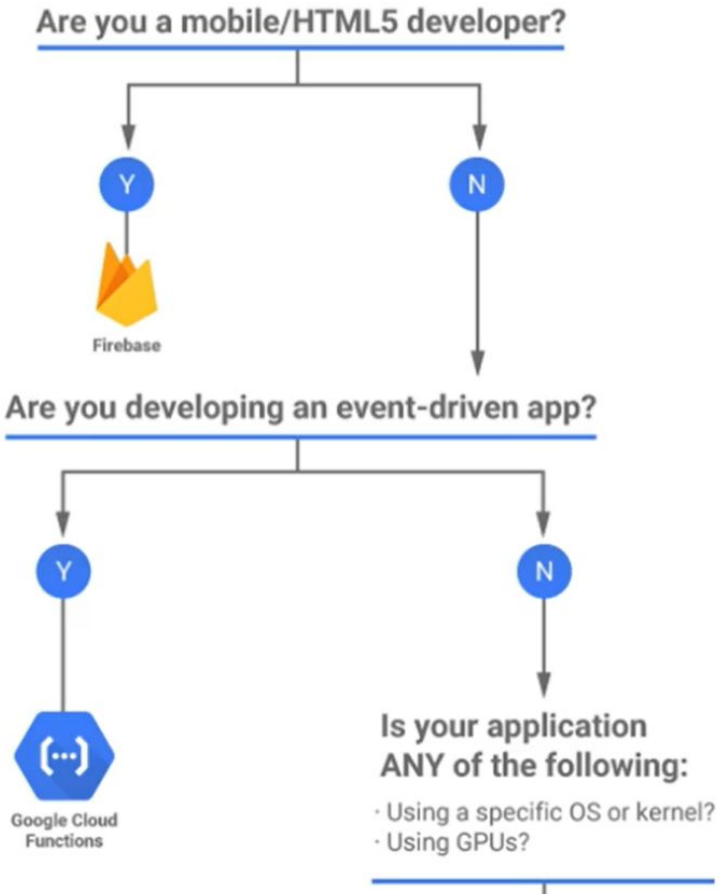
Data Flows - Processing: Cloud Functions (GCF)



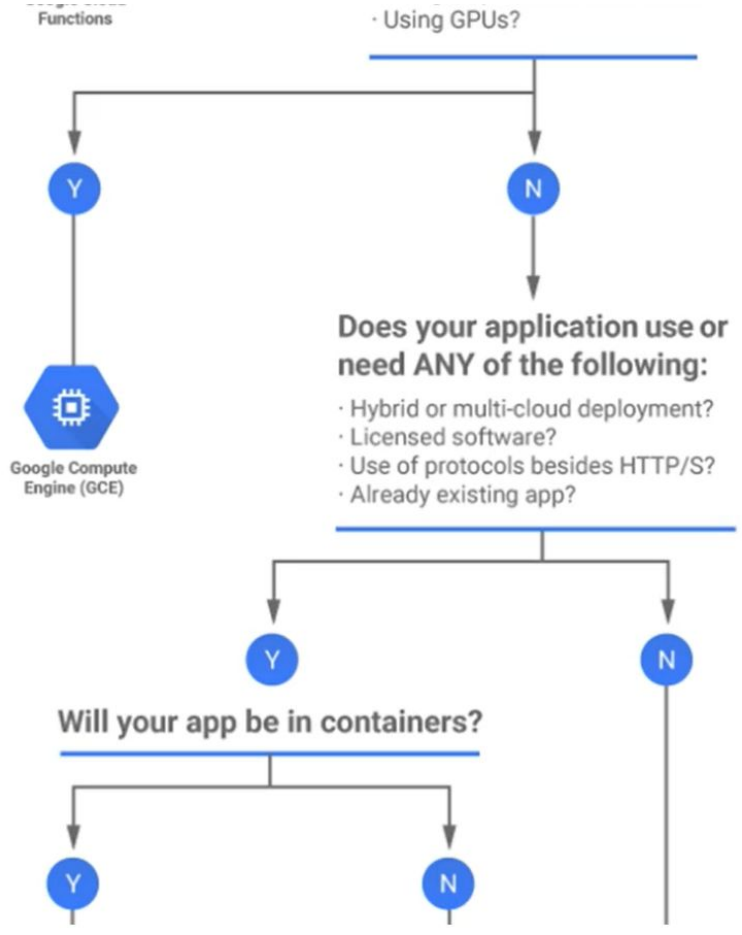
Zonal Regional Multi-Regional Global

- Runs code in response to an event — Node.js, Python, Java, Go
- Functions as a Service (FaaS), “Serverless”
- Pay for CPU and RAM assigned to function, per 100ms (min. 100ms)
- Each function automatically gets an HTTP endpoint
- Can be triggered by GCS objects, Pub/Sub messages, etc.
- Massively scalable (horizontally) — Runs many copies when needed
- Often used for chatbots, message processors, IoT, automation, etc.

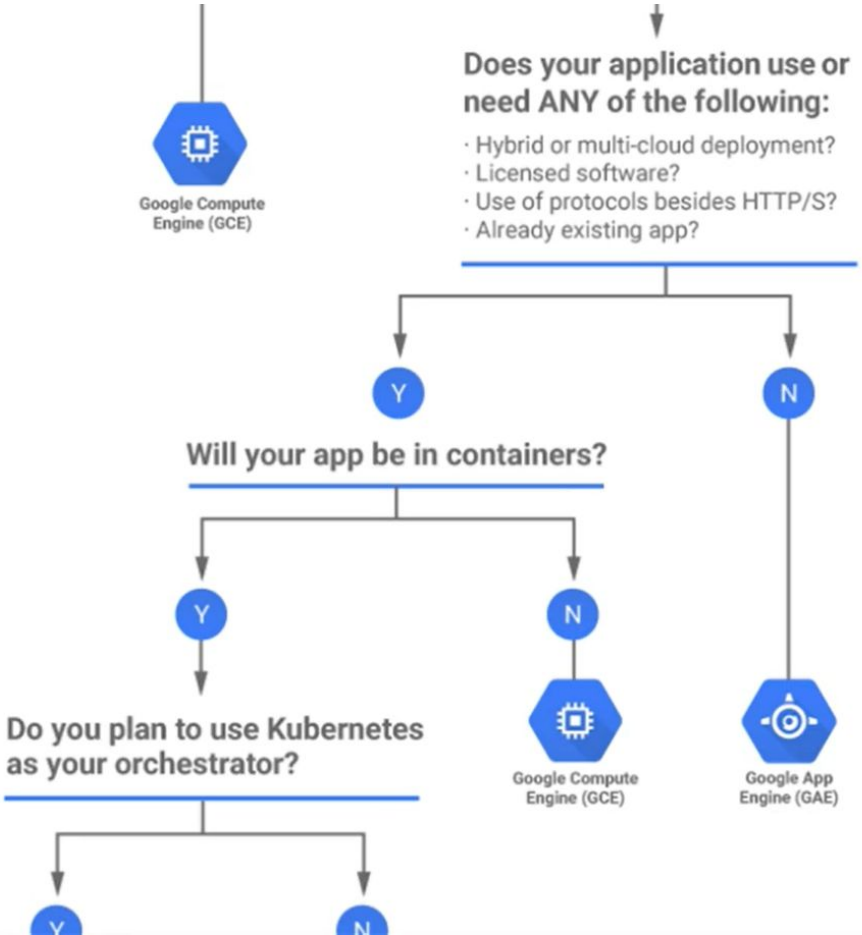
Choosing compute: Flowchart



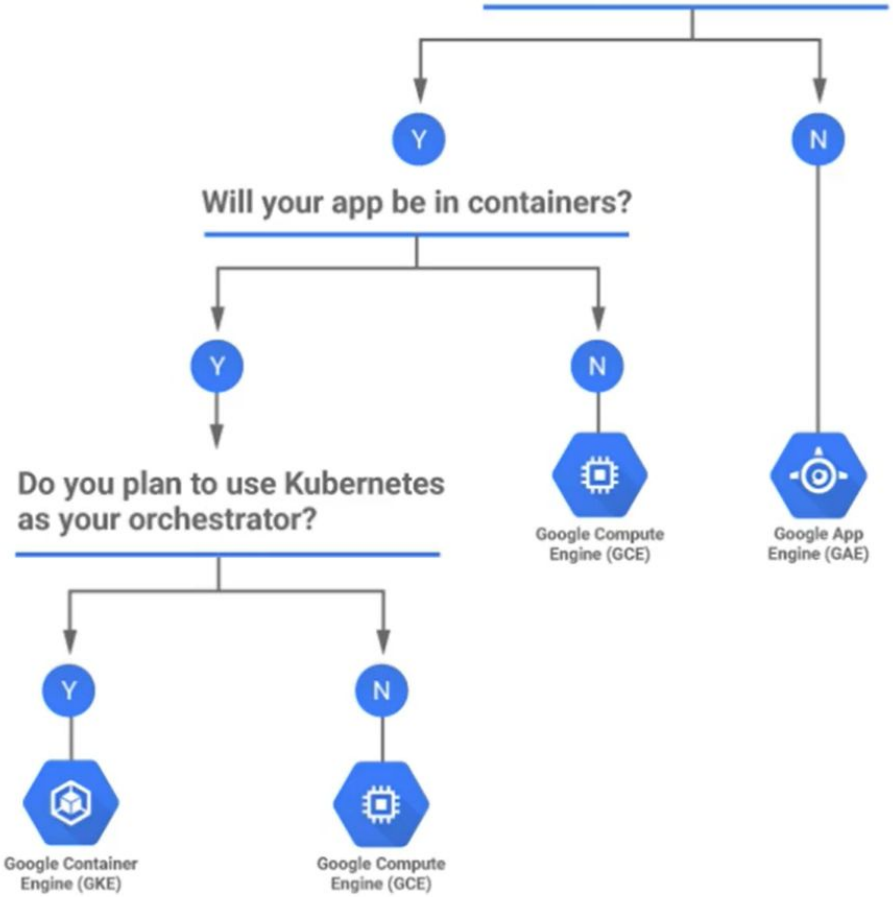
Choosing compute: Flowchart



Choosing compute: Flowchart



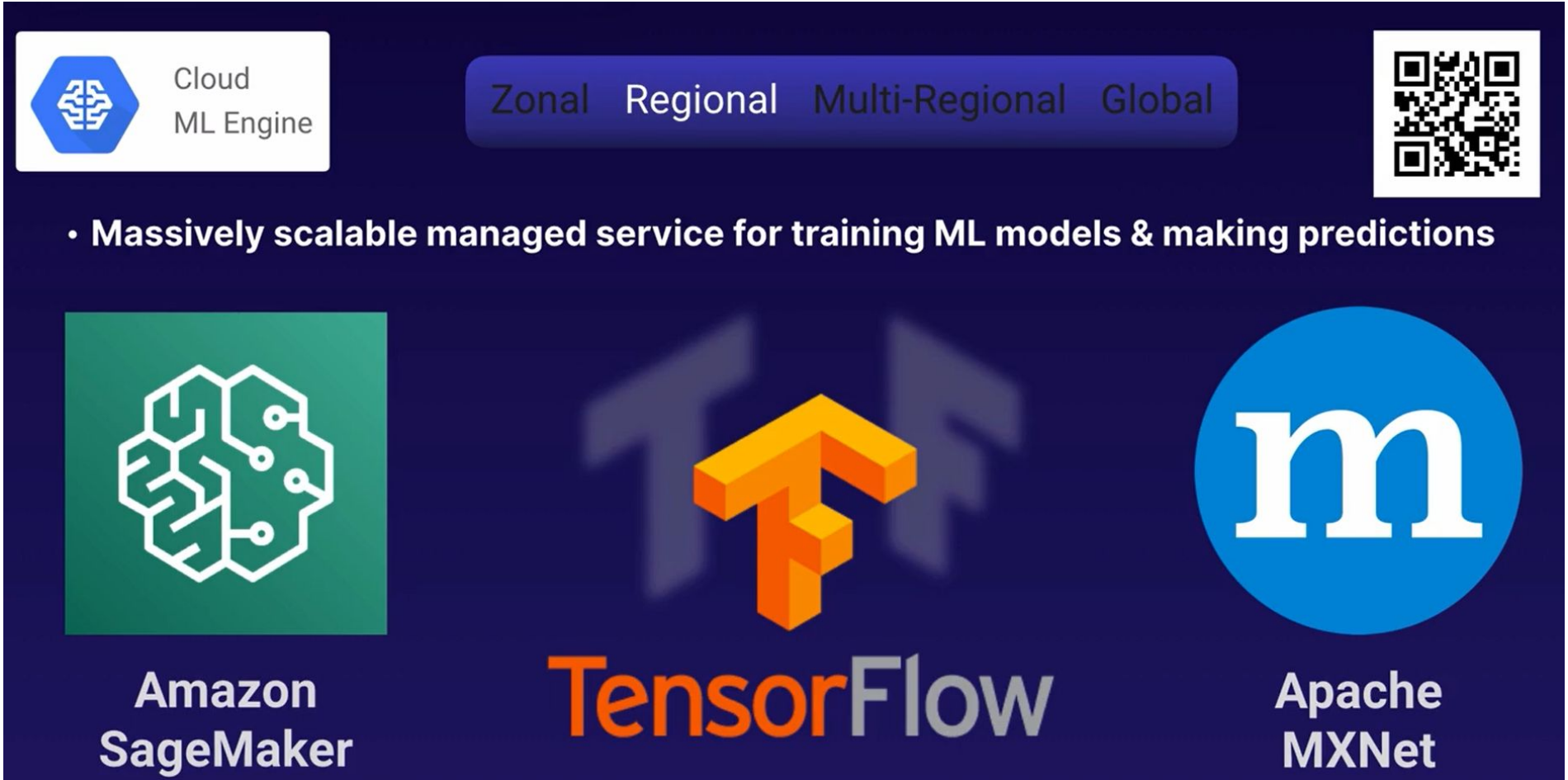
Choosing compute: Flowchart



Data Flows - Processing: Autoscaling Compute Comparison

Autoscaling	Minimum	Min for HA	Mechanism	Speed
Compute Engine	1 instance	2 instances	MIG alone / MIG + CLB	Moderate / Better
GKE Pods	1 pod	2-pod deployment	Load-based internal K8s	Decent
GKE Node Pool	1 instance	3 instances	K8s pod back-pressure	Slow
App Engine Flex	1 container	2 containers	Request-based	"Gradually"
App Engine Standard	Zero	Zero	Request-based	"Sudden and extreme spikes of traffic" OK
Cloud Functions	Zero	Zero	Request-based	Very Fast

Data Flows - Processing: Machine Learning and AI

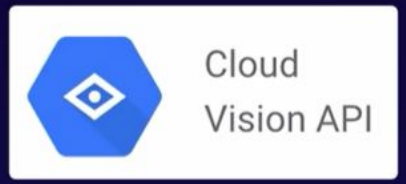


Data Flows - Processing: Machine Learning and AI

Zonal Regional Multi-Regional Global

- Massively scalable managed service for training ML models & making predictions
- Enables apps/devs to use TensorFlow on datasets of any size; endless use cases
- Integrates: GCS/BQ (storage), Cloud Datalab (dev), Cloud Dataflow (preprocessing)
- Supports online & batch predictions, prioritizing latency (online) & job time (batch)
- Or download models & make predictions anywhere: desktop, mobile, own servers
- HyperTune automatically tunes model hyperparameters to avoid manual tweaking
- Training: Pay per hour depending on chosen cluster capabilities (ML training units)
- Prediction: Pay per provisioned node-hour plus by prediction request volume made

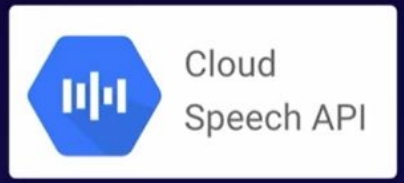
Data Flows - Processing: Machine Learning and AI



Zonal Regional Multi-Regional Global

- Classifies images into categories, detects objects/faces, & finds/reads printed text
- Pre-trained ML model to analyze images and discover their contents
- Classifies into thousands of categories (e.g., "sailboat", "lion", "Eiffel Tower")
- Upload images or point to ones stored in GCS
- Pay per image, based on detection features requested
  - Higher price for OCR of full documents and finding similar images on the web
  - Some features are prices together: Labels + SafeSearch, ImgProps + Cropping
  - Other features priced individually: Text, Faces, Landmarks, Logos

Data Flows - Processing: Machine Learning and AI



Zonal Regional Multi-Regional Global

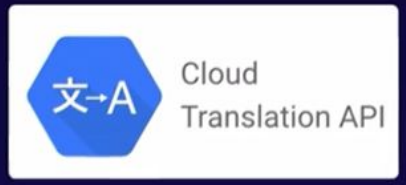
- Automatic Speech Recognition (ASR) to turn spoken word audio files into text
- Pre-trained ML model for recognizing speech in 110+ languages/variants
- Accepts pre-recorded or real-time audio, & can stream results back in real-time
- Enables voice command-and-control and transcribing user microphone dictations
- Handles noisy source audio
- Optionally filters inappropriate content in some languages
- Accepts contextual hints: words and names that will likely be spoken
- Pay per 15 seconds of audio processed

Data Flows - Processing: Machine Learning and AI

Zonal Regional Multi-Regional Global

- Analyzes text for sentiment, intent, & content classification, and extracts info
- Pre-trained ML model for understanding what text means, so you can act on it
- Excellent with Speech API (audio), Vision API (OCR), & Translation API (or built-ins)
- Syntax analysis extracts tokens/sentences, parts of speech & dependency trees
- Entity analysis finds people, places, things, etc., labels them, & links to Wikipedia
- Analysis for sentiment (overall) and entity sentiment detect +/- feelings & strength
- Content classification puts each document into one of 700+ predefined categories
- Charged per request of 1000 characters, depending on analysis types requested

Data Flows - Processing: Machine Learning and AI



Zonal Regional Multi-Regional Global

- Translate text among 100+ languages; optionally auto-detects source language
- Pre-trained ML model for recognizing and translating semantics, not just syntax
- Can let people support multi-regional clients in non-native languages, 2-way
- Combine with Speech, Vision, & Natural Language APIs for powerful workflows
- Send plain text or HTML and receive translation in kind
- Pay per character processed for translation
- Also pay per character for language auto-detection

Data Flows - Processing: Machine Learning and AI

Zonal Regional Multi-Regional Global

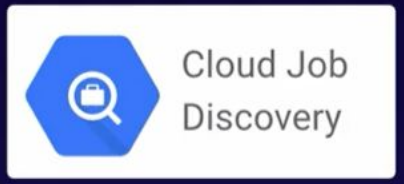
- Build conversational interfaces for websites, mobile apps, messaging, IoT devices
- Pre-trained ML model and service for accepting, parsing, lexing input & responding
- Enables useful chatbots and other natural user interactions with your custom code
- Train it to identify custom entity types by providing a small dataset of examples
- Or choose from 30+ pre-built agents (e.g. car, currency, dates) as starting template
- Supports many different languages and platforms/devices
- Free plan has unlimited text interactions and capped voice interactions
- Paid plan is unlimited but charges per request: more for voice, less for text

Data Flows - Processing: Machine Learning and AI

Zonal Regional Multi-Regional Global

- Annotates videos in GCS (or directly uploaded) with info about what they contain
- Pre-trained ML model for video scene analysis and subject identification
- Enables you to search a video catalog the same way you search text documents
- “Specify a region where processing will take place (for regulatory compliance)”
- *Label Detection*: Detect entities within the video, such as "dog", "flower" or "car"
- *Shot Change Detection*: Detect scene changes within the video
- *SafeSearch Detection*: Detect adult content within the video
- Pay per minute of video processed, depending on requested detection modes

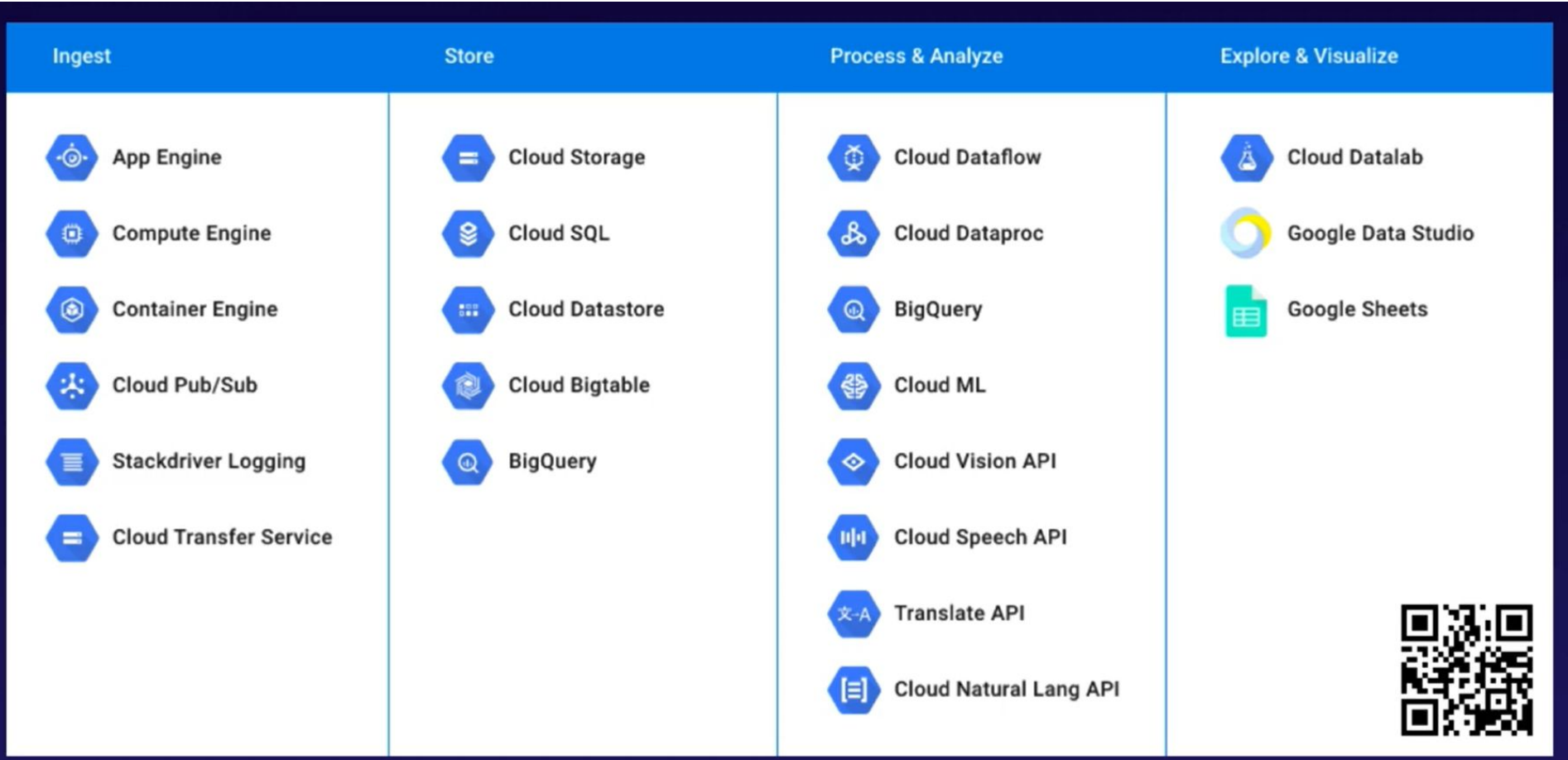
Data Flows - Processing: Machine Learning and AI



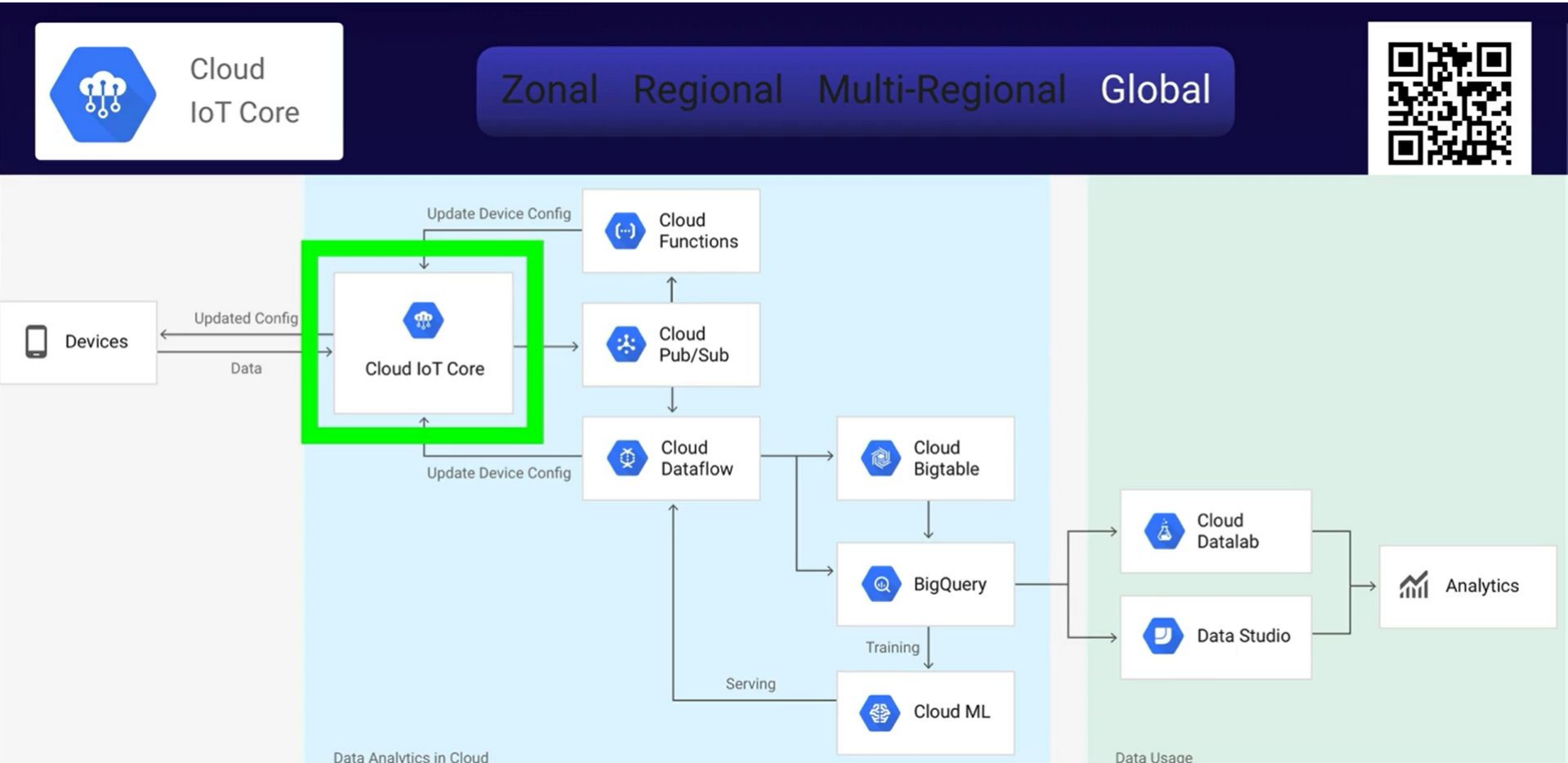
Zonal Regional Multi-Regional Global

- Helps career sites, company job boards, etc. to improve engagement & conversion
- Pre-trained ML model to help job seekers search job posting databases
- `Most job sites rely on keyword search to retrieve content which often omits relevant jobs and overwhelms the job seeker with irrelevant jobs. For example, a keyword search with any spelling error returns 0 results, and a keyword search for “dental assistant” returns any "assistant" role that offers dental benefits.’
- Integrates with many job/hiring systems
- Lots of features, such as commute distance and recognizing abbreviations/jargon
- “Show me jobs with a 30 minute commute on public transportation from my home”

Data Flows - Processing: Big Data Lifecycle



Data Flows - Processing: Big Data and IoT



Data Flows - Processing: Big Data and IoT

Zonal Regional Multi-Regional Global

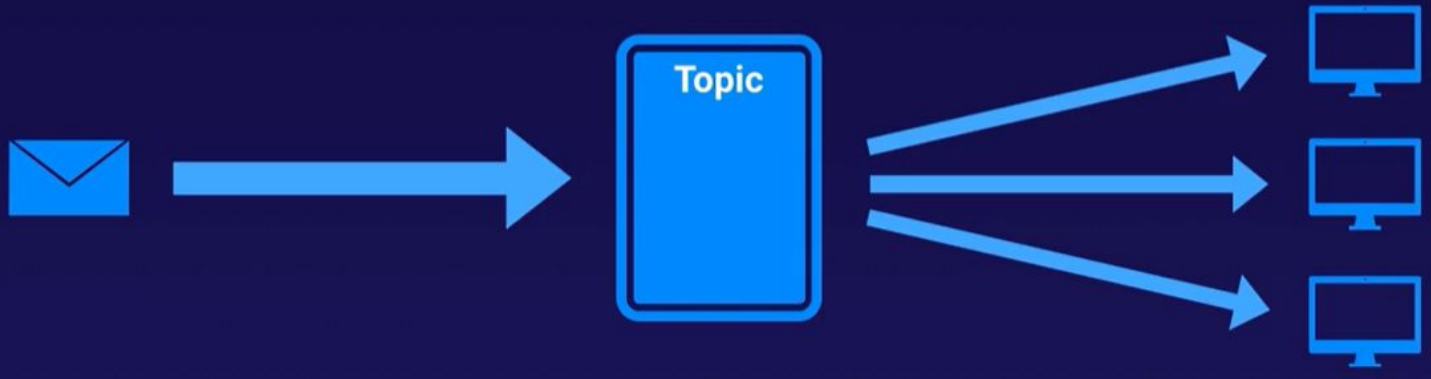
- Fully-managed service to connect, manage, and ingest data from devices globally
- Device Manager handles device identity, authentication, config, & control
- Protocol Bridge publishes incoming telemetry to Cloud Pub/Sub for processing
- Connect securely using IoT industry-standard MQTT or HTTPS protocols
- CA signed certificates can be used to verify device ownership on first connect
- Two-way device communication enables configuration & firmware updates
- Device shadows enable querying & making control changes while devices offline

Data Flows - Processing: Big Data and IoT

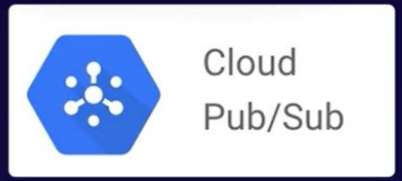
Cloud Pub/Sub

Zonal Regional Multi-Regional Global

- Infinitely-scalable at-least-once messaging for ingestion, decoupling, etc.



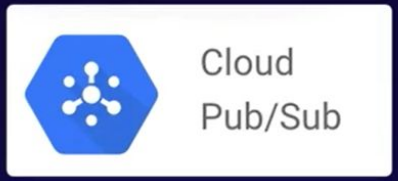
Data Flows - Processing: Big Data and IoT



Zonal Regional Multi-Regional Global

- Infinitely-scalable at-least-once messaging for ingestion, decoupling, etc.

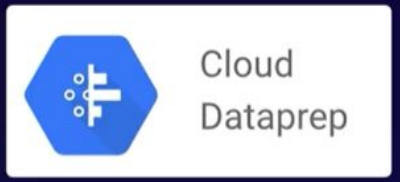
Data Flows - Processing: Big Data and IoT



Zonal Regional Multi-Regional Global

- Infinitely-scalable at-least-once messaging for ingestion, decoupling, etc.
- “Global by default: Publish… and consume from anywhere, with consistent latency.”
- Messages can be up to 10MB and undelivered ones stored for 7 days—but no DLQ
- Push mode delivers to HTTPS endpoint & succeeds on HTTP success status code
  - “Slow-start” algorithm ramps up on success and backs off & retries, on failures
- Pull mode delivers messages to requesting clients and waits for ACK to delete
  - Lets clients set rate of consumption, and supports batching and long-polling
- Pay for data volume
  - Min 1KB per publish/push/pull request (not by message)

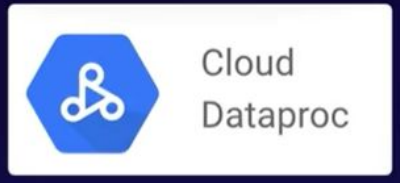
Data Flows - Processing: Big Data and IoT



Zonal Regional Multi-Regional Global

- Visually explore, clean, and prepare data for analysis without running servers
- "Data Wrangling" (i.e. "ad-hoc ETL") for business analysts, not IT pros
  - Who might otherwise spend 80% of their time cleaning data
- Managed version of Trifacta Wrangler—and managed by Trifacta, not Google
- Source data from GCS, BQ, or file upload—formatted in CSV, JSON, or relational
- Automatically detects schemas, datatypes, possible joins, and various anomalies
- Pay for underlying Dataflow job, plus management overhead charge

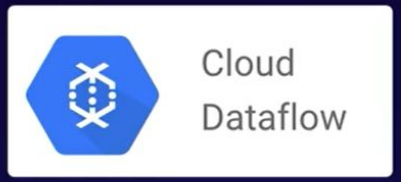
Data Flows - Processing: Big Data and IoT



Zonal Regional Multi-Regional Global

- Batch MapReduce processing via configurable, managed Spark & Hadoop clusters
- Handles being told to scale (adding or removing nodes) even while running jobs
- Integrated with Cloud Storage, BigQuery, Bigtable, and some Stackdriver services
- “Image versioning” switches between versions of Spark, Hadoop, & other tools
- Pay directly for underlying GCE servers used in the cluster—optionally preemptible
- Pay a Cloud Dataproc management fee per vCPU-hour in the cluster
- Best for moving *existing* Spark/Hadoop setups to GCP
  - Prefer Cloud Dataflow for new data processing pipelines — “Go with the flow”

Data Flows - Processing: Big Data and IoT

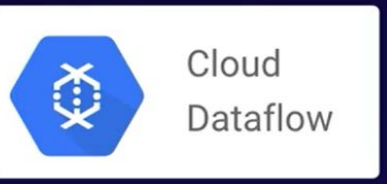


Zonal Regional Multi-Regional Global

- Smartly-autoscaled & fully-managed batch or stream MapReduce-like processing
- Released as open-source Apache Beam

Amazon EMR

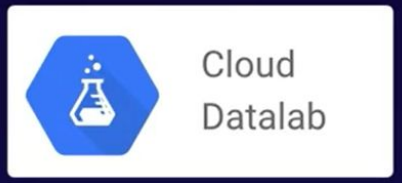
Data Flows - Processing: Big Data and IoT



Zonal Regional Multi-Regional Global

- Smartly-autoscaled & fully-managed batch or stream MapReduce-like processing
- Released as open-source Apache Beam
- Autoscales & dynamically redistributes lagging work, mid-job, to optimize run time
- Integrated with Cloud Pub/Sub, Datastore, BQ, Bigtable, Cloud ML, Stackdriver, etc.
- Dataflow Shuffle service for batch offloads Shuffle ops from workers for big gains
- Effectively pay for underlying worker GCE via consolidated charges
  - Pay per second for vCPUs, RAM GBs PD/PD-SSD (more for streaming)
- Dataflow Shuffle charged for time per GB used

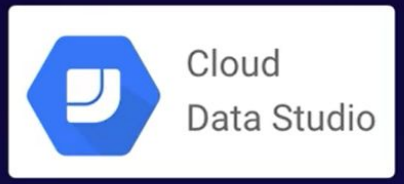
Data Flows - Processing: Big Data and IoT



Zonal Regional Multi-Regional Global

- Interactive tool for data exploration, analysis, visualization and machine learning
- Uses Jupyter Notebook
  - "[A]n open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more."
- Supports iterative development of data analysis algorithms in Python/SQL/~JS
- Pay for GCE/GAE instance hosting and storing (on PD) your notebooks
- Pay for any other resources accessed (e.g. BigQuery)

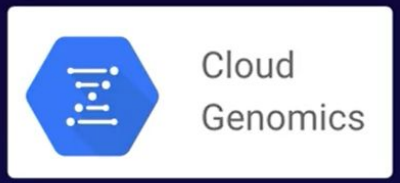
Data Flows - Processing: Big Data and IoT



Zonal Regional Multi-Regional Global

- Big Data Visualization tool for dashboards and reporting
- Meaningful data stories/presentations enable better business decision making
- Data sources include BigQuery, Cloud SQL, other MySQL, Google Sheets, Google Analytics, Analytics 360, AdWords, DoubleClick, & YouTube channels
- Visualizations include time series, bar charts, pie charts, tables, heat maps, geo maps, scorecards, scatter charts, bullet charts, & area charts
- Templates for quick start; customization options for impactful finish
- Familiar G Suite sharing and real-time collaboration
- Pay only for services accessed

Data Flows - Processing: Big Data and IoT



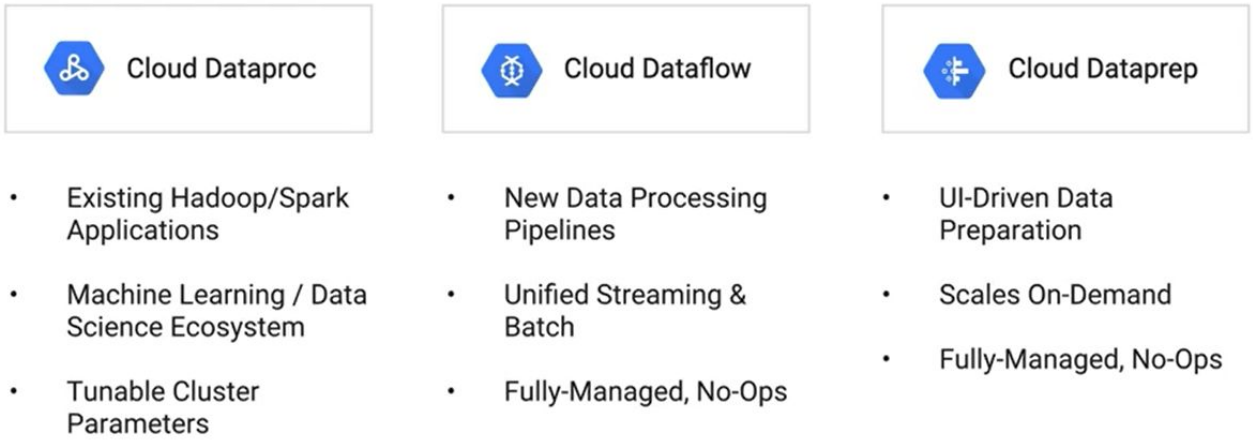
Zonal Regional Multi-Regional Global

- Store and process genomes and related experiments
- Query complete genomic information of large research projects in seconds
- Process many genomes and experiments in parallel
- Open industry standards (e.g. from Global Alliance for Genomics and Health)
- Supports “Requester Pays” sharing

Data Flows - Processing: Dataproc or Dataflow

Processing large-scale data

Large-scale data processing typically involves reading data from source systems such as Cloud Storage, Cloud Bigtable, or Cloud SQL, and then conducting complex normalizations or aggregations of that data. In many cases, the data is too large to fit on a single machine so frameworks are used to manage distributed compute clusters and to provide software tools that aid processing.



Data Flows - Processing: Dataproc or Dataflow

Recommended Workloads

WORKLOADS	CLOUD DATAPROC	CLOUD DATAFLOW
Stream processing (ETL)		✓
Batch processing (ETL)	✓	✓
Iterative processing and notebooks	✓	
Machine learning with Spark ML	✓	
Preprocessing for machine learning		✓ (with Cloud ML Engine)

Hands-on lab

Data Flows - Storage: Local SSD

Local SSD

Zonal Regional Multi-Regional Global

- Very fast 375GB solid state drives physically attached to the server

EC2 Instance Store Vols.

Direct-Attached Storage (DAS)

Data Flows - Storage: Local SSD

Zonal Regional Multi-Regional Global

- Very fast 375GB solid state drives physically attached to the server
- Can stripe across eight of them (3TB) for *even* better performance
- DATA WILL BE LOST whenever the instance shuts down
  - But can survive a Live Migration
- Like all data at rest, always encrypted

Data Flows - Storage: Persistent Disk (PD)

Zonal Regional Multi-Regional Global

- Flexible, *block-based* network-attached storage; boot disk for every GCE instance
- Perf scales with volume size; max way below Local SSD, but still plenty fast
- Persistent disks *persist*, and are replicated (zone or region) for durability
- Can resize while in use (up to 64TB), but will need file system update within VM
- Snapshots (and machine images) add even more capability and flexibility
  - "Magical": Pay for incremental ($ and time), but use/delete like full backups
- *Not* file-based NAS, but can mount to multiple instances if *all* are read-only
- Pay for GB/mo provisioned depending on perf. class; plus snapshot GB/mo used

Data Flows - Storage: Cloud Filestore

Cloud Filestore

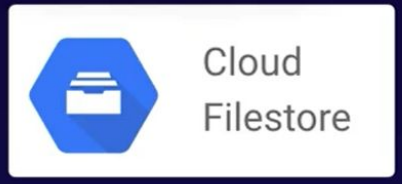
Zonal Regional Multi-Regional Global

- **Fully-managed file-based storage**

Elastic File System (EFS)

Network-Attached Storage (NAS)

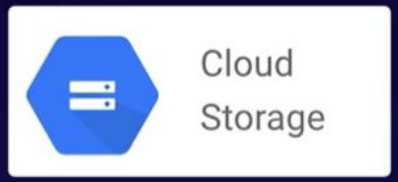
Data Flows - Storage: Cloud Filestore



Zonal Regional Multi-Regional Global

- Fully-managed file-based storage
- “Predictably fast performance for your file-based workloads”
- Accessible to GCE and GKE through your VPC, via NFSv3 protocol
- Primary use case is application migration to cloud (“lift and shift”)
- Fully manages file *serving*, but *not* backups
- Pay for provisioned TBs in “Standard” (slow) or “Premium” (fast) mode
- Minimum provisioned capacity of 1TB (Standard) or 2.5TB (Premium)

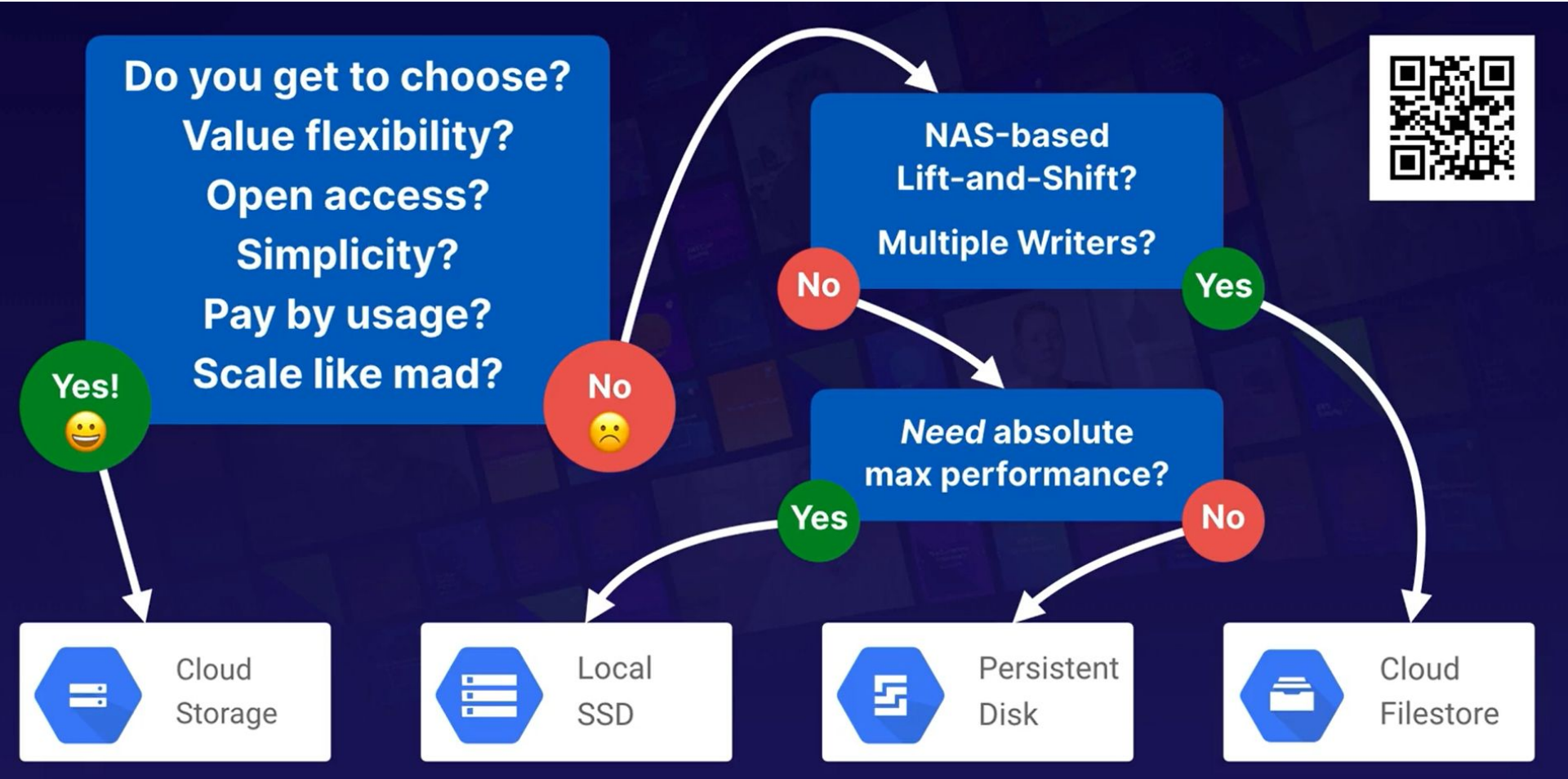
Data Flows - Storage: Cloud Storage



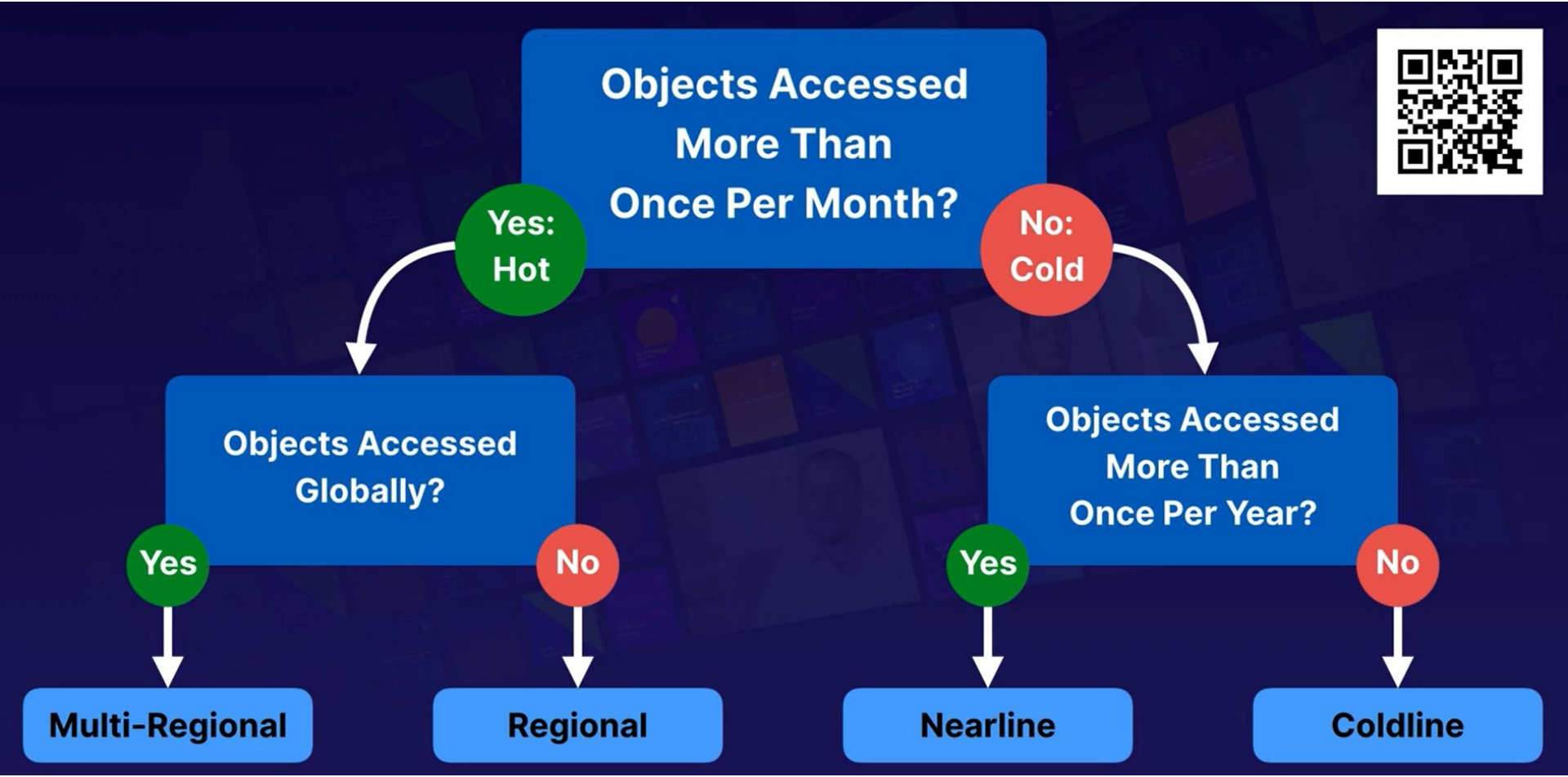
Zonal Regional Multi-Regional Global

- Infinitely scalable, fully-managed, versioned, and highly-durable object storage
- Designed for 99.999999999% (“eleven nines”) durability
- Strongly consistent (even for overwrite PUTs and DELETEs)
- Integrated site hosting and CDN functionality
- Lifecycle transitions across classes: Multi-Regional, Regional, Nearline, Coldline
  - Diffs in cost & availability (99.95%, 99.9%, 99%, 99%), not latency (no thaw delay)
- All classes have same API, so can use gsutil and gcsfuse (but beware)
- Pay for data operations & GB-months stored by class
- Nearline/Coldline: Also pay for GBs retrieved—plus early deletion fee, if < 30/90 days

Choosing your storage: Flowchart



Choosing a cloud storage class: Flowchart



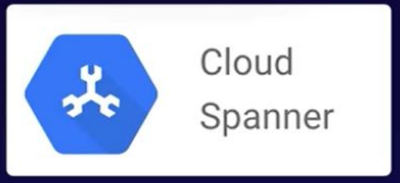
Databases: CloudSQL

Databases: CloudSQL

Zonal Regional Multi-Regional Global

- Fully-managed and reliable MySQL and PostgreSQL databases
- Supports automatic replication, backup, failover, etc.
- Scaling is manual (both vertically and horizontally)
- Effectively pay for underlying GCE instances and PDs
  - Plus some baked-in service fees

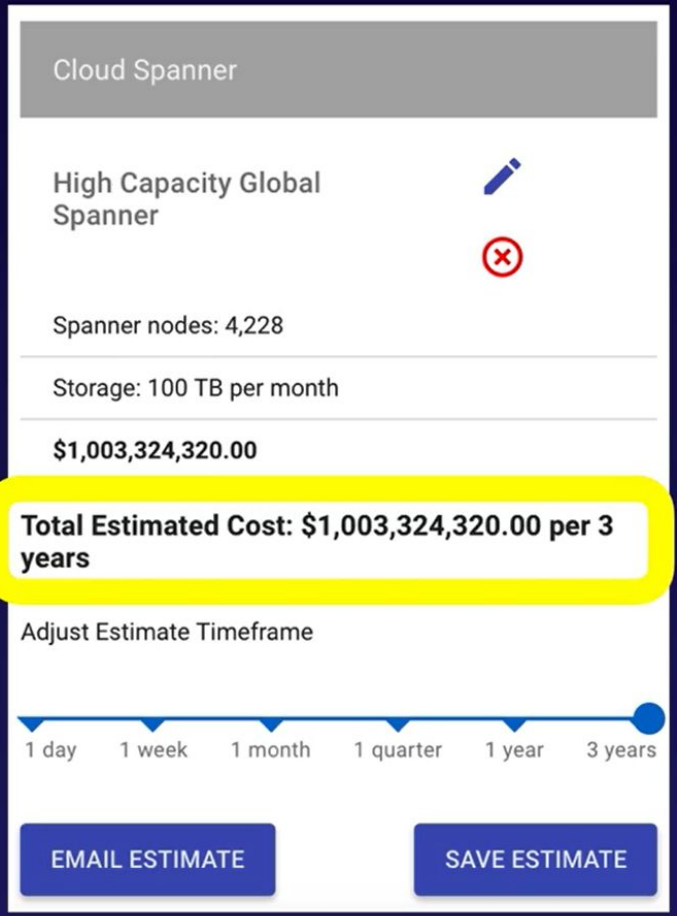
Databases: Cloud Spanner



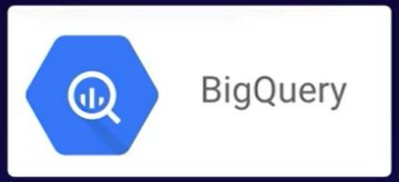
Zonal Regional Multi-Regional Global

- "The first horizontally scalable, strongly consistent, relational database service"
  - "From 1 to hundreds or thousands of nodes"
  - "A minimum of 3 nodes is recommended for production environments."
- Chooses Consistency and Partition-Tolerance (CP of CAP theorem)
- But still *high* Availability: SLA has 99.999% SLO (five nines) for multi-region
  - Nothing is actually 100%, really
  - *Not* based on fail-over
- Pay for provisioned node time (by region/multi-region) plus used storage-time

Databases: Cloud Spanner



Databases: BigQuery



Zonal Regional Multi-Regional Global

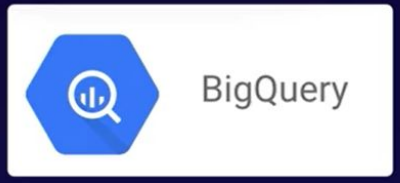
- Serverless column-store data warehouse for analytics using SQL
- Scales internally, so it “can scan TB in seconds and PB in minutes”

Amazon Redshift

Amazon Athena

Presto

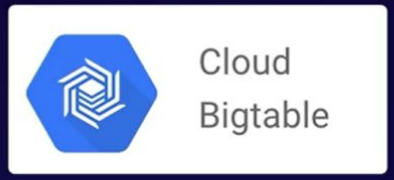
Databases: BigQuery



Zonal Regional Multi-Regional Global

- Serverless column-store data warehouse for analytics using SQL
- Scales internally, so it “can scan TB in seconds and PB in minutes”
- Pay for GBs actually considered (scanned) during queries
  - Attempts to reuse cached results, which are free
- Pay for data stored (GB-months)
  - Relatively inexpensive
  - Even cheaper when table not modified for 90 days (reading still fine)
- Pay for GBs added via streaming inserts

Databases: Cloud Bigtable



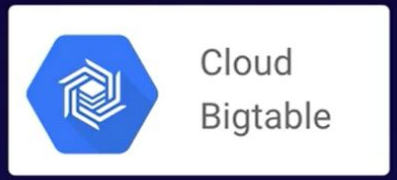
Zonal Regional Multi-Regional Global

- Low latency & high throughput NoSQL DB for large operational & analytical apps

DynamoDB

Cassandra

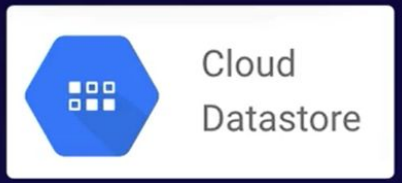
Databases: Cloud Bigtable



Zonal Regional Multi-Regional Global

- Low latency & high throughput NoSQL DB for large operational & analytical apps
- Supports open-source HBase API
- Integrates with Hadoop, Dataflow, Dataproc
- Scales seamlessly and unlimitedly
  - Storage autoscales
  - Processing nodes must be scaled manually
- Pay for processing node hours

Databases: Cloud Datastore



Zonal Regional Multi-Regional Global

- Managed & autoscaled NoSQL DB with indexes, queries, and ACID trans. support
- NoSQL, so queries can get complicated
  - No joins or aggregates and must line up with indexes
  - NOT, OR, and NOT EQUALS (<>, !=) operations not natively supported
- Automatic “built-in” indexes for simple filtering and sorting (ASC, DESC)
- Manual “composite” indexes for more complicated, but beware them “exploding”
- Pay for GB-months of storage used (including indexes)
- Pay for IO operations (deletes, reads, writes) performed (i.e. no pre-provisioning)

Databases: Firebase

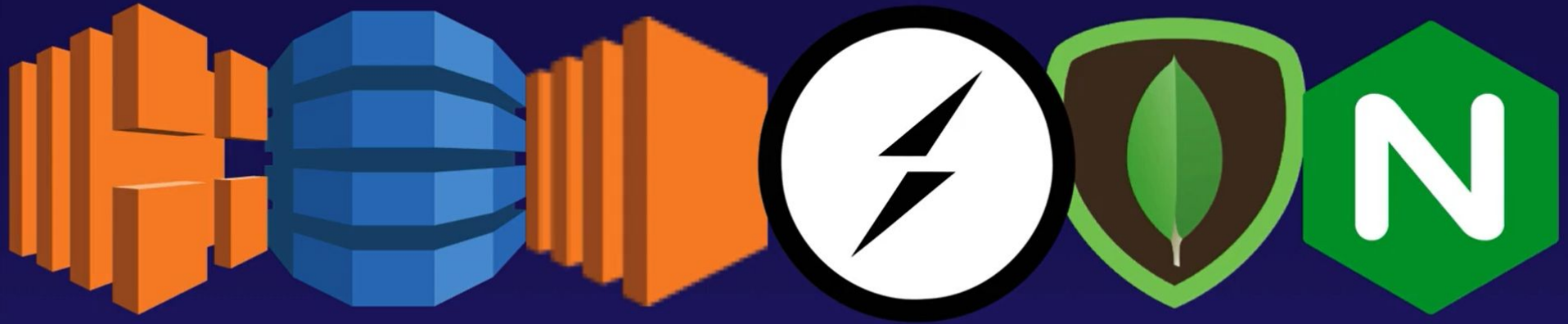
Firebase Realtime DB

Zonal Regional Multi-Regional Global

Cloud Firestore

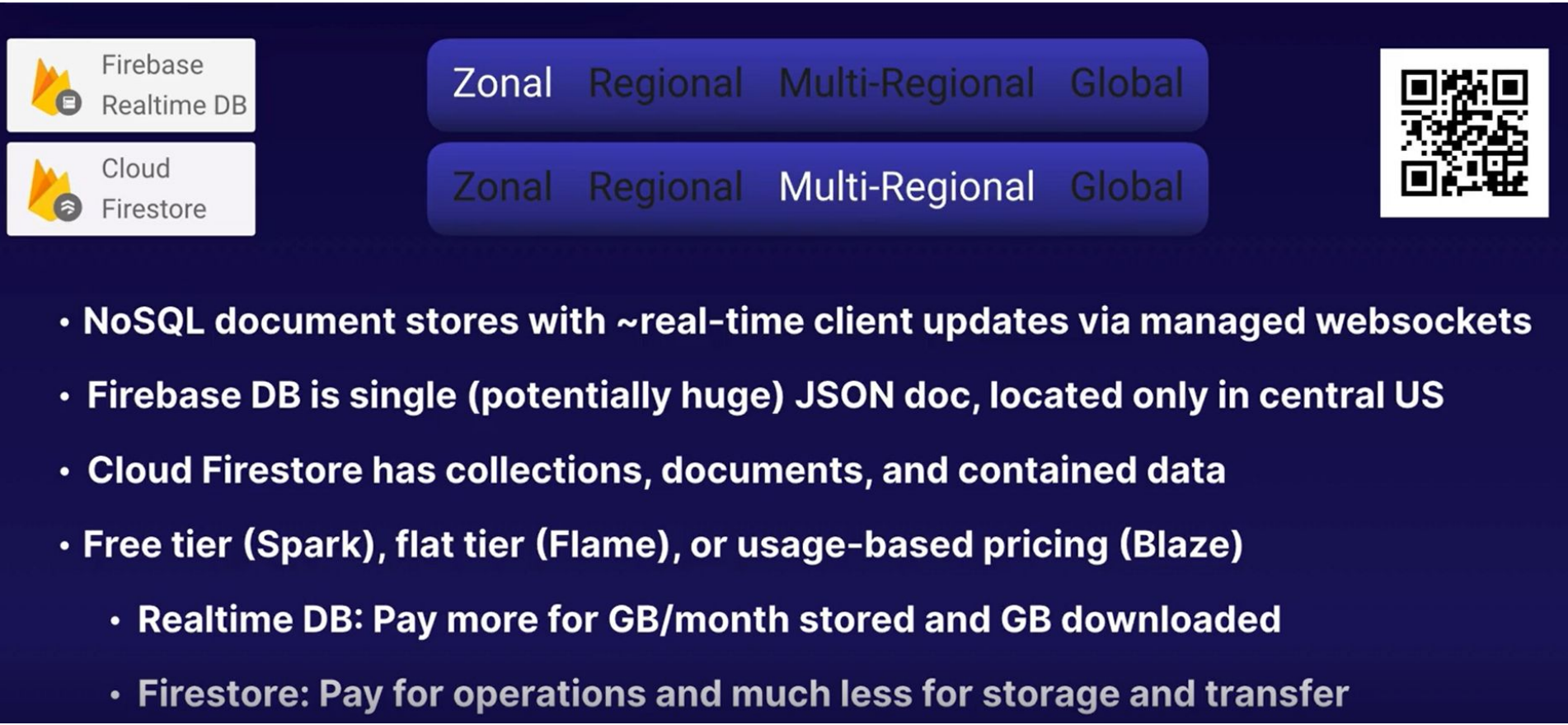
Zonal Regional Multi-Regional Global

- NoSQL document stores with ~real-time client updates via managed websockets



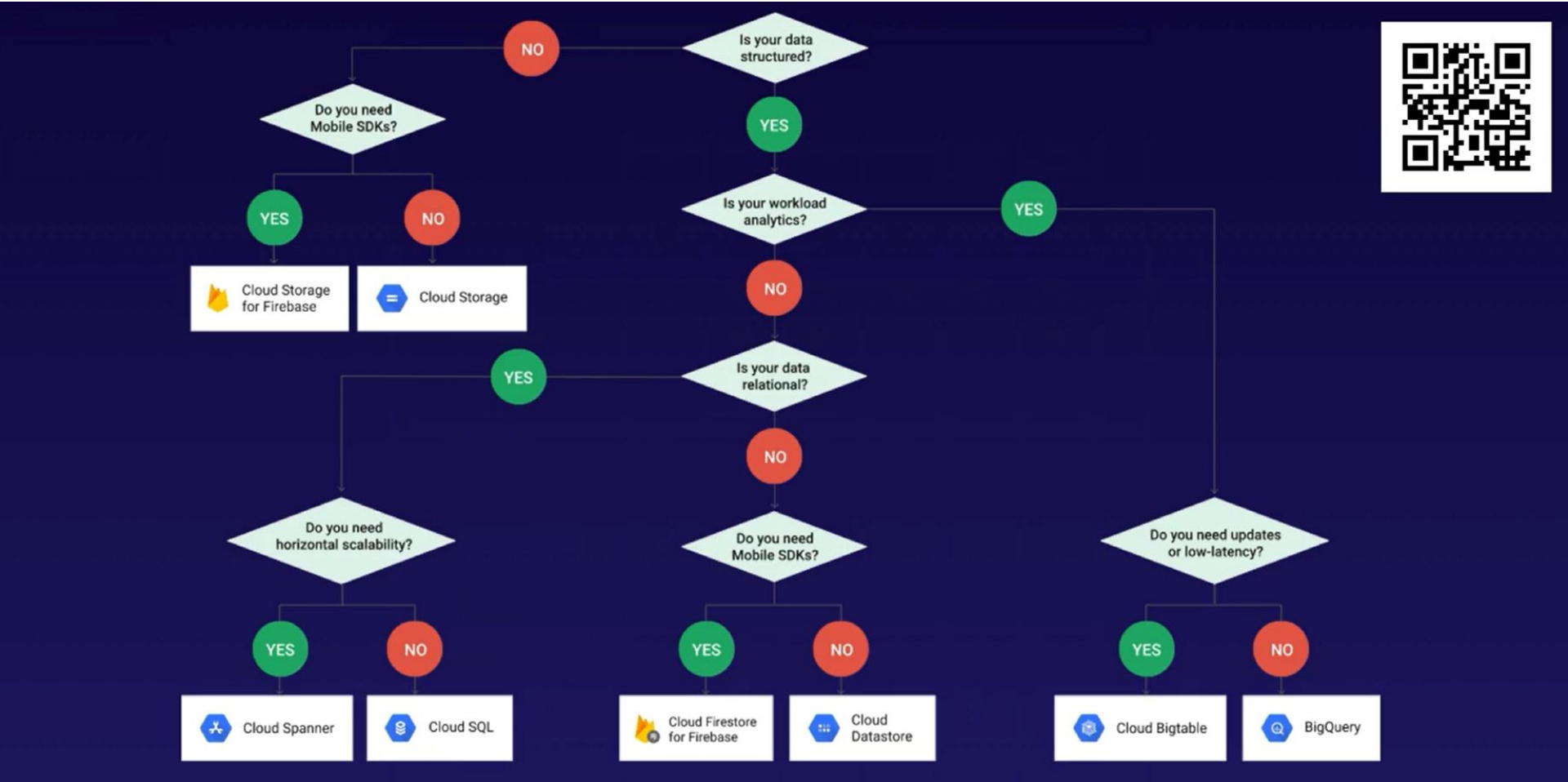
ELB DynamoDB EC2 Socket.io MongoDB NGINX

Databases: Cloud Spanner

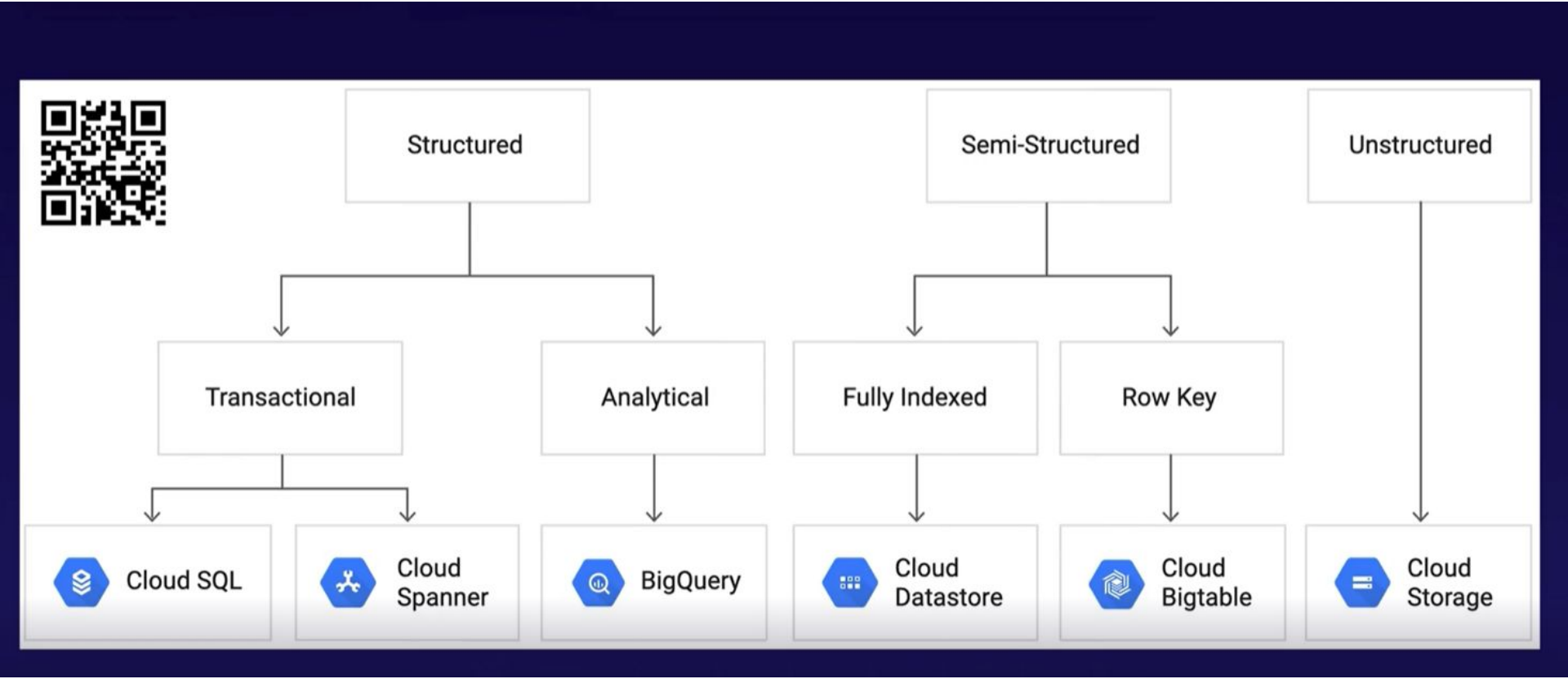


- NoSQL document stores with ~real-time client updates via managed websockets
- Firebase DB is single (potentially huge) JSON doc, located only in central US
- Cloud Firestore has collections, documents, and contained data
- Free tier (Spark), flat tier (Flame), or usage-based pricing (Blaze)
  - Realtime DB: Pay more for GB/month stored and GB downloaded
  - Firestore: Pay for operations and much less for storage and transfer

Choosing a database: Flowchart



Choosing a database: Data-first apparoach



Hands-on lab

Network Layer Design

Data transfer

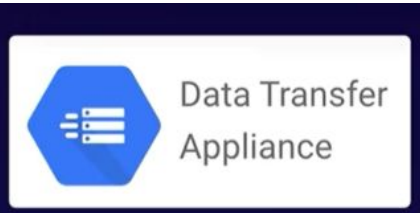
Data Transfer: Data Transfer Appliance

Data Transfer Appliance

- Rackable, high-capacity storage server to physically ship data to GCS

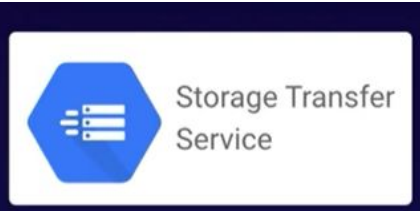
AWS Snowball

Data Transfer: Data Transfer Appliance



- Rackable, high-capacity storage server to physically ship data to GCS
- Ingest only; not a way to avoid egress charges
- 100TB or 480TB versions
- 480TB/week is faster than a saturated 6Gbps link

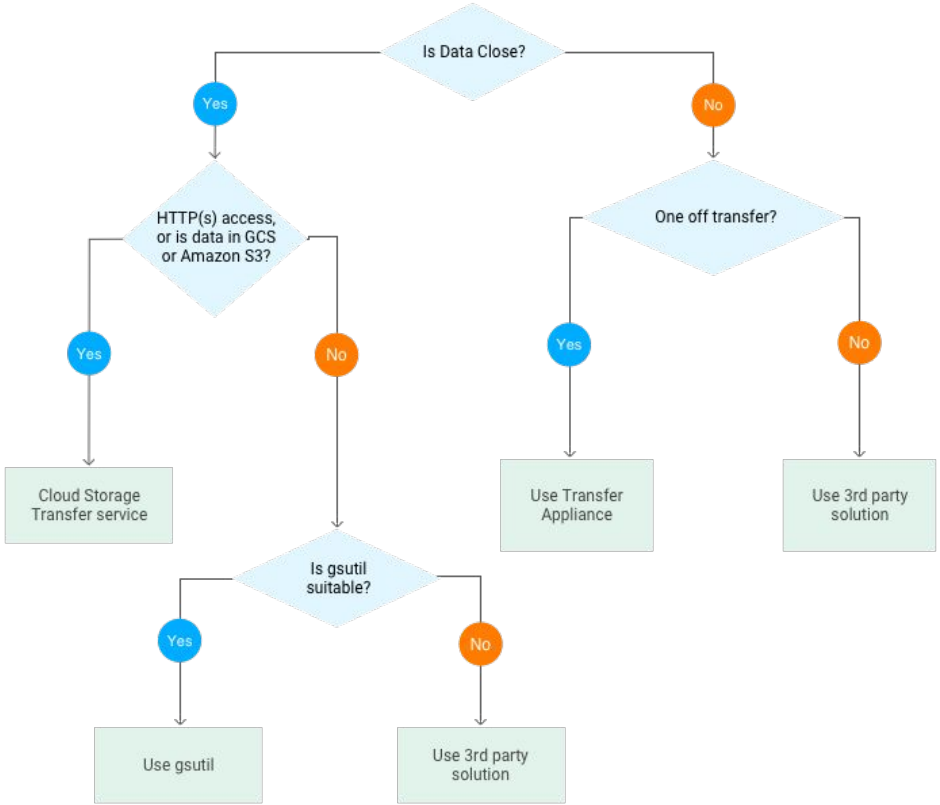
Data Transfer: Storage Transfer Service



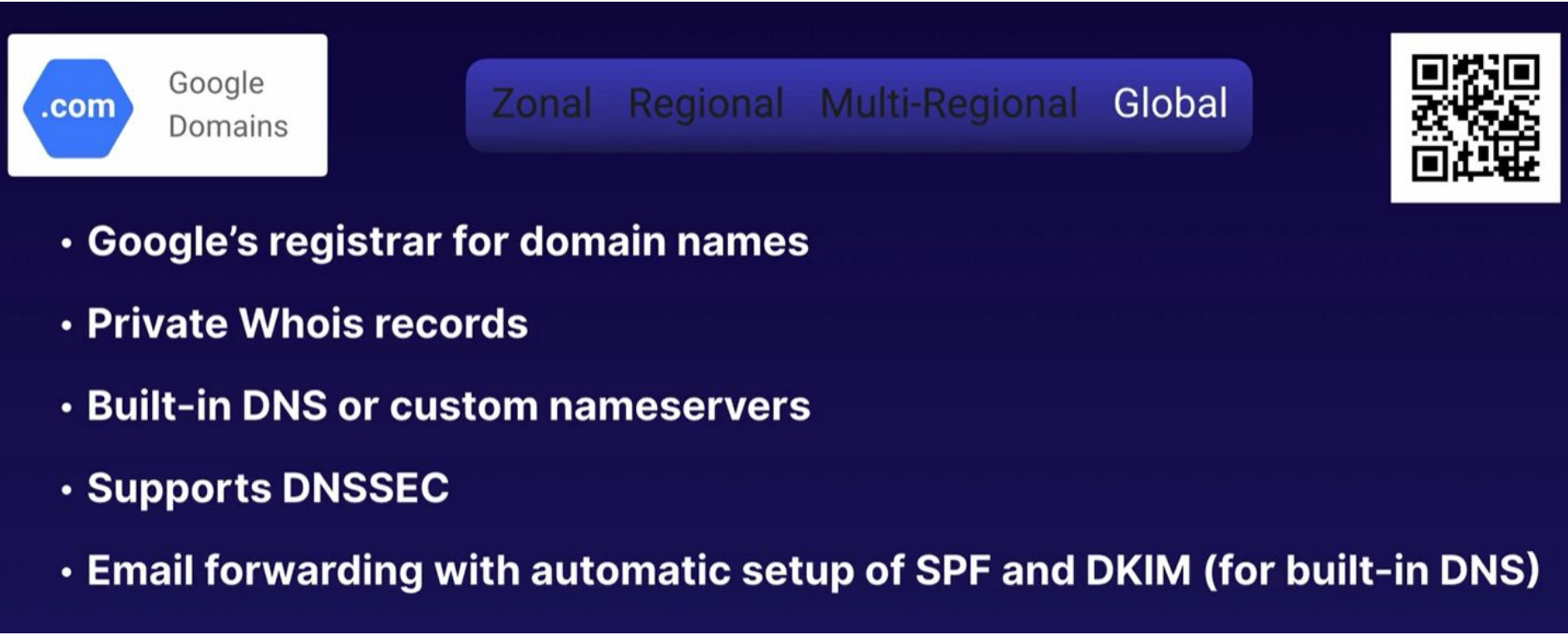
Zonal Regional Multi-Regional Global

- Copies objects for you, so you don't need to set up a machine to do it
- Destination is always GCS bucket
- Source can be S3, HTTP/HTTPS endpoint, or another GCS bucket
- One-time or scheduled recurring transfers
- Free to use, but you pay for its actions

Data transfer options : Flowchart

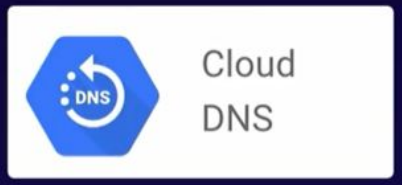


Data Transfer: Google Domains



Data Transfer: Cloud DNS

Data Transfer: Cloud DNS



Zonal Regional Multi-Regional Global

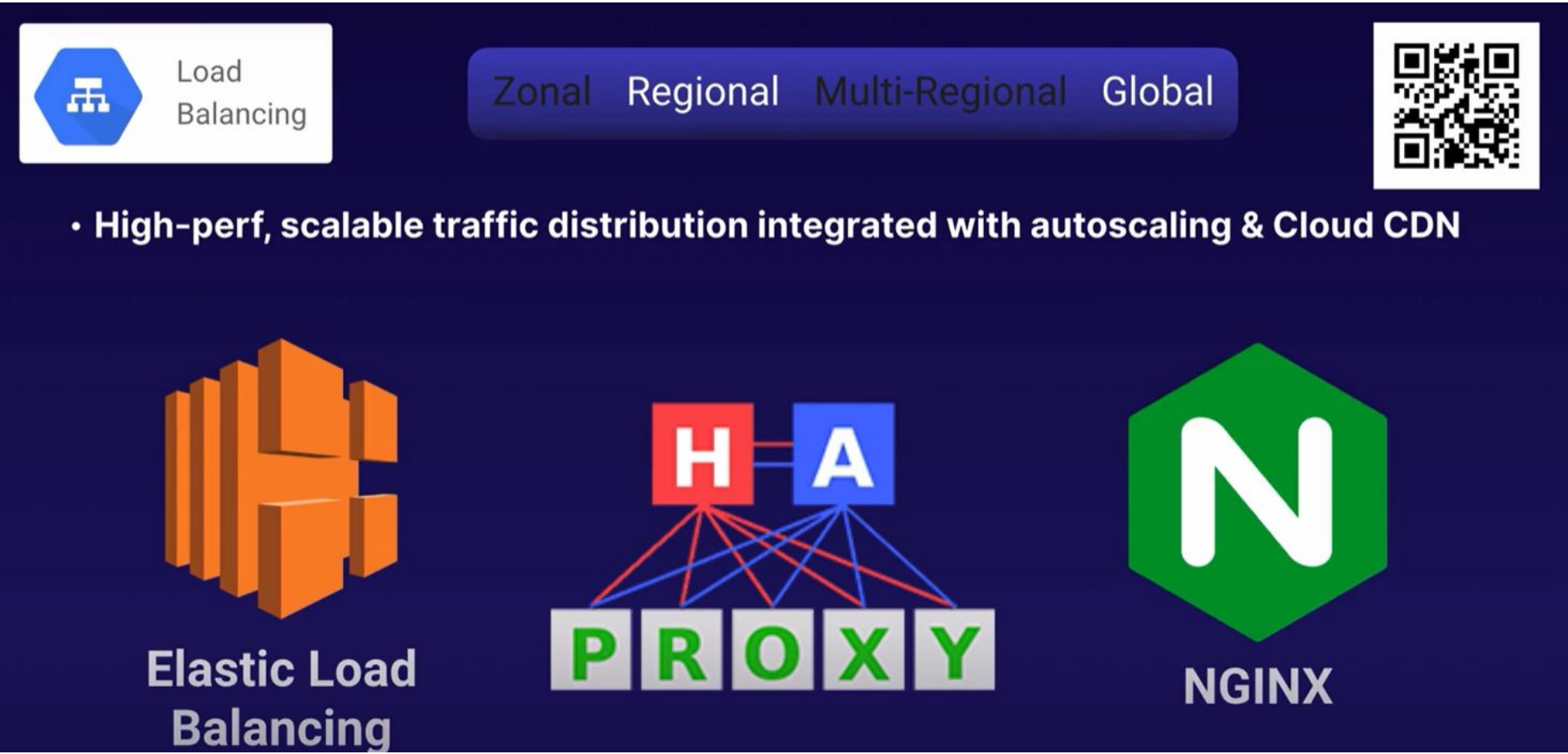
- Scalable, reliable, & managed authoritative Domain Name System (DNS) service
- 100% uptime guarantee
- Public and private managed zones
- Low latency globally
- Supports DNSSEC
- Manage via UI, CLI, or API
- Pay fixed fee per managed zone to store and distribute DNS records
- Pay for DNS lookups (i.e. usage)

Data Transfer: Static IP

Zonal Regional Multi-Regional Global

- Reserve static IP addresses in projects and assign them to resources
- Regional IPs used for GCE instances & Network Load Balancers
- Global IPs used for global load balancers:
  - HTTP(S), SSL proxy, and TCP proxy
  - “Anycast IP” simplifies DNS
- Pay for reserved IPs that are not in use, to discourage wasting them

Data Transfer: Load Balancing



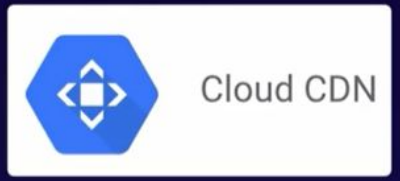
Data Transfer: Load Balancing

Zonal Regional Multi-Regional Global

- High-perf, scalable traffic distribution integrated with autoscaling & Cloud CDN
- SDN naturally handles spikes without any prewarming; no instances or devices
- Regional Network Load Balancer: health checks, round robin, session affinity
  - Forwarding rules based on IP, protocol (e.g. TCP, UDP), and (optionally) port
- Global load balancers w/ multi-region failover for HTTP(S), SSL proxy, & TCP proxy
  - Prioritize low-latency connection to region near user, then gently fail over in bits
  - Reacts quickly (unlike DNS) to changes in users, traffic, network, health, etc.
- Pay by making ingress traffic billable (cheaper than egress) plus hourly per rule

Data Transfer: Cloud CDN

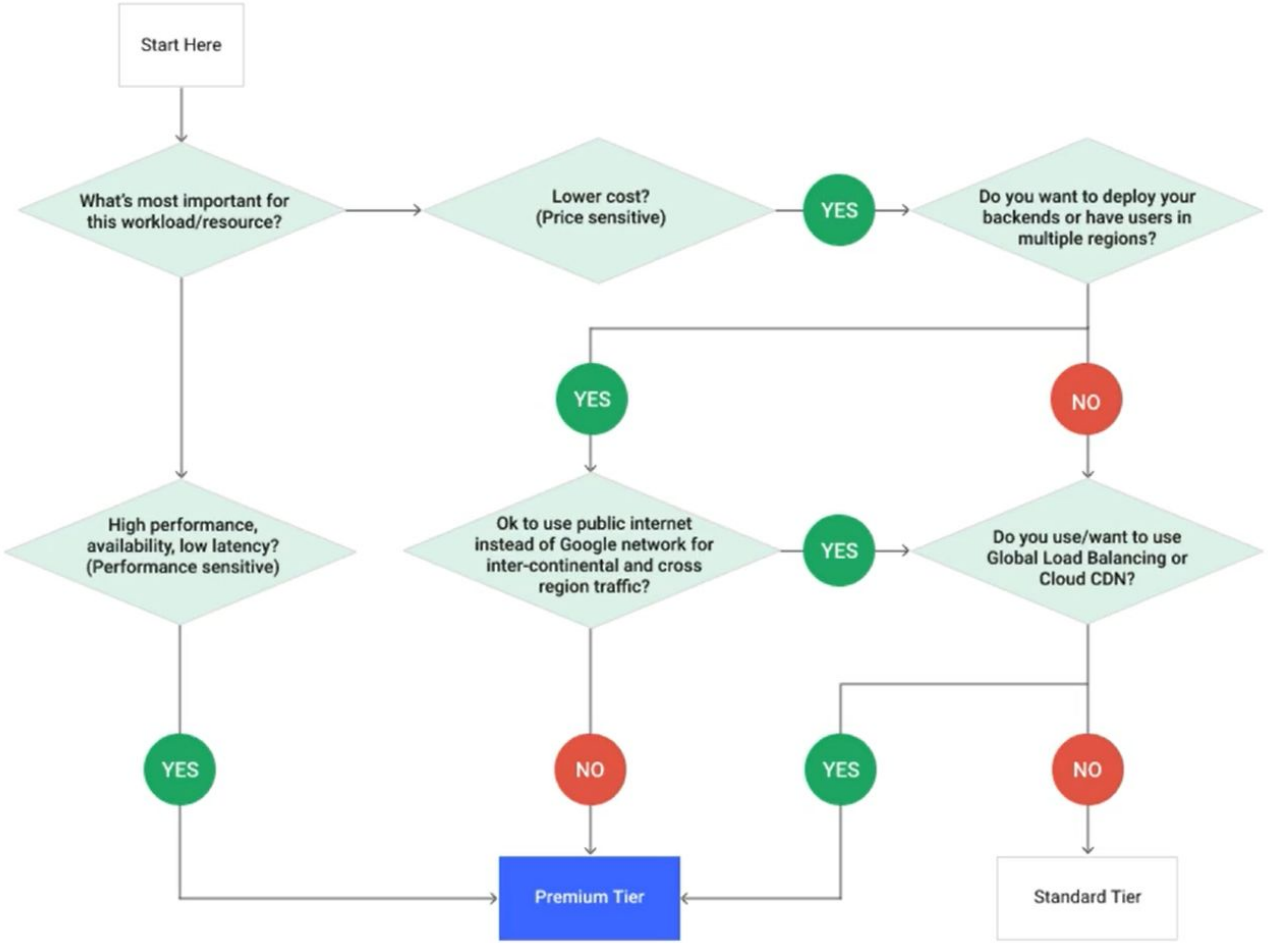
Data Transfer: Cloud CDN



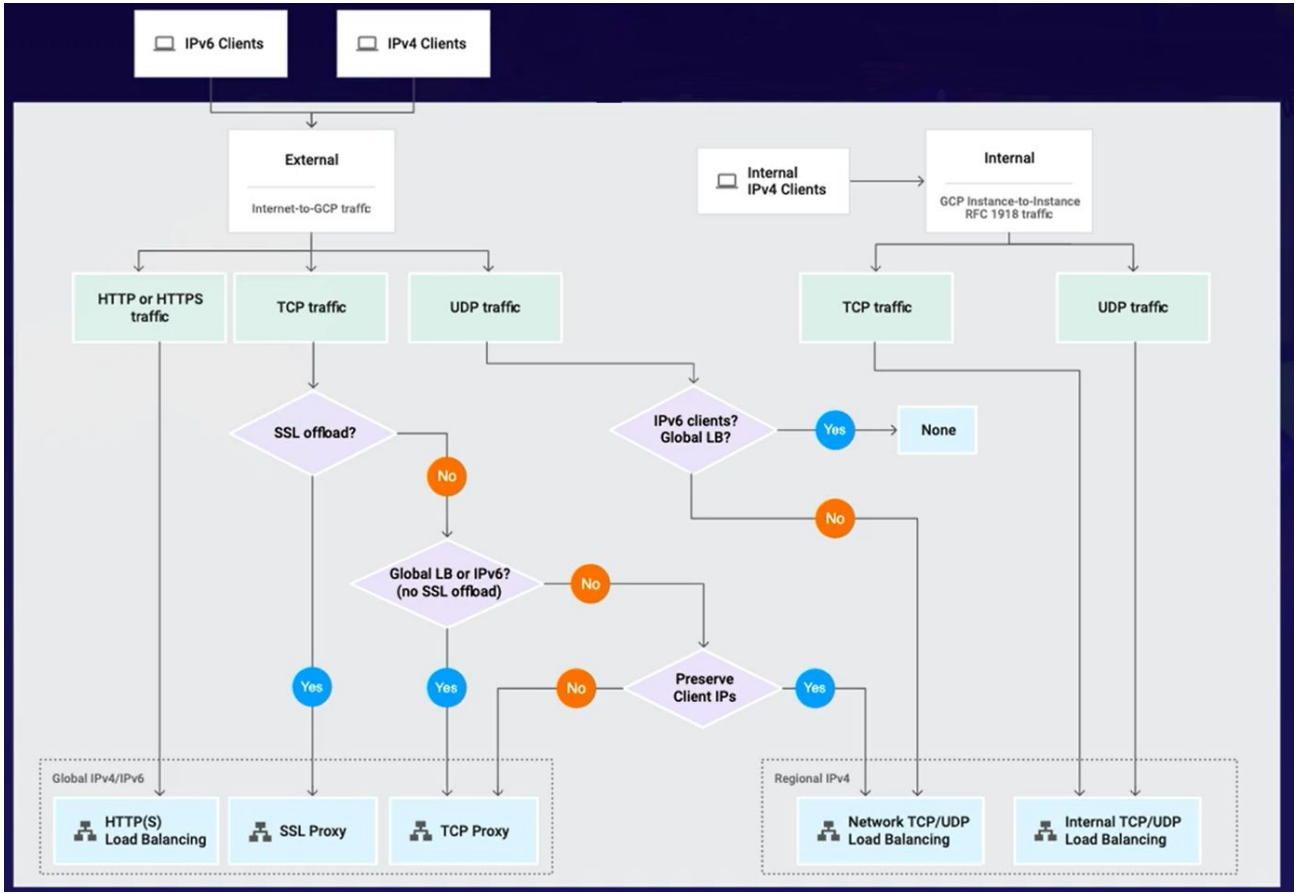
Zonal Regional Multi-Regional Global

- Low-latency content delivery based on HTTP(S) CLB & integrated w/ GCE & GCS
- Supports HTTP/2 and HTTPS, but no custom origins (GCP only)
- Simple checkbox on HTTP(S) Load Balancer config turns this on
- On cache miss, pay origin→POP “cache fill” egress charges (cheaper for in-region)
- Always pay POP→client egress charges, depending on location
- Pay for HTTP(S) request volume
- Pay per cache invalidation request (not per resource invalidated)
- Origin costs (e.g. CLB, GCS) can be much lower because cache hits reduce load

Network Tier Choice : Flowchart



Load Balancer Options : Flowchart



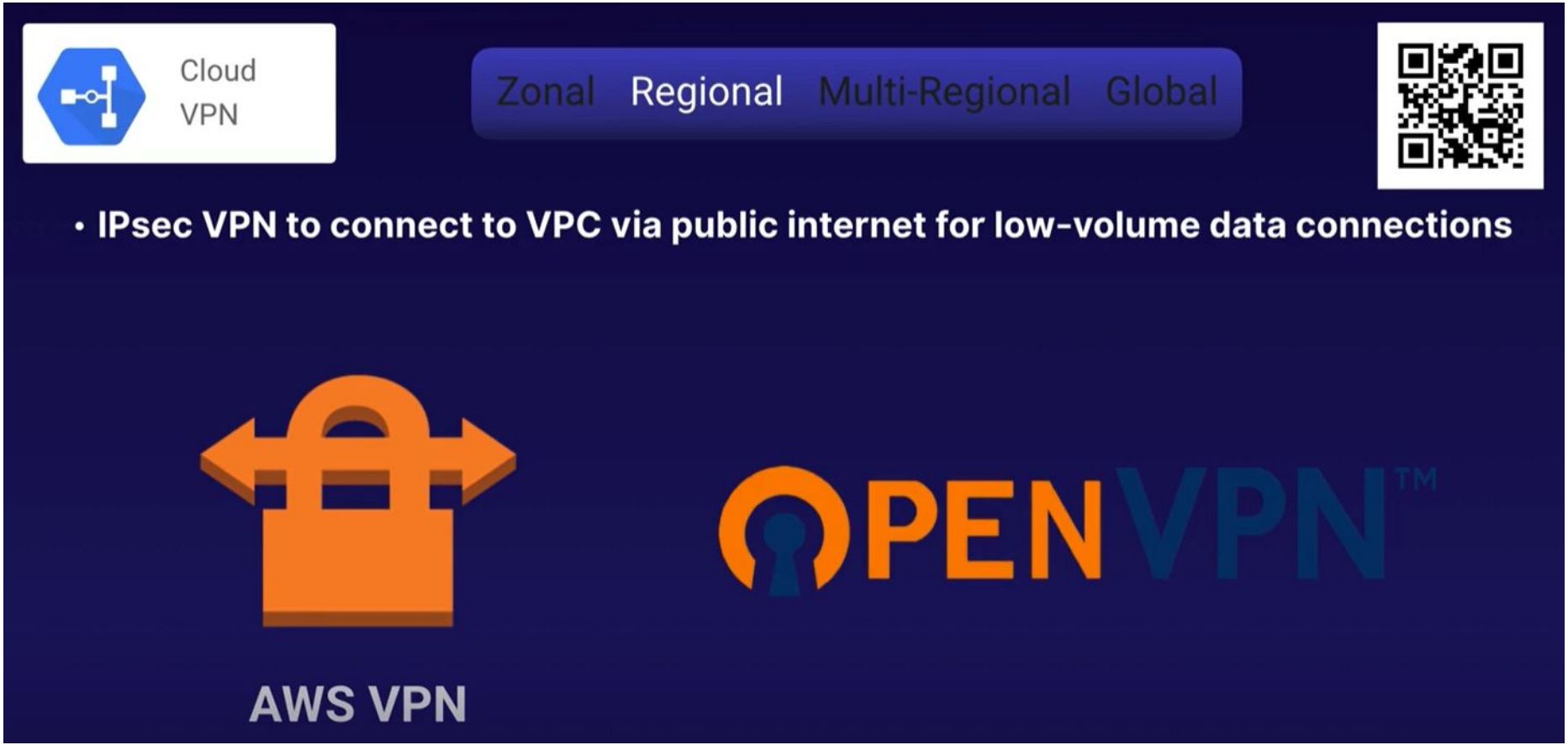
Internal Networking: VPC

Zonal Regional Multi-Regional Global

- Global IPv4 unicast Software-Defined Network (SDN) for GCP resources
- Automatic mode is easy; custom mode gives control
- Configure subnets (each with a private IP range), routes, firewalls, VPNs, BGP, etc.
- VPC is global and subnets are regional (not zonal!)
- Can be shared across multiple projects in same org and peered with other VPCs
- Can enable private (internal IP) access to some GCP services (e.g. BQ, GCS)
- Free to configure VPC (container)
- Pay to use certain services (e.g. VPN) and for network egress

Internal Networking: Cloud Interconnect

Internal Networking: Cloud VPN

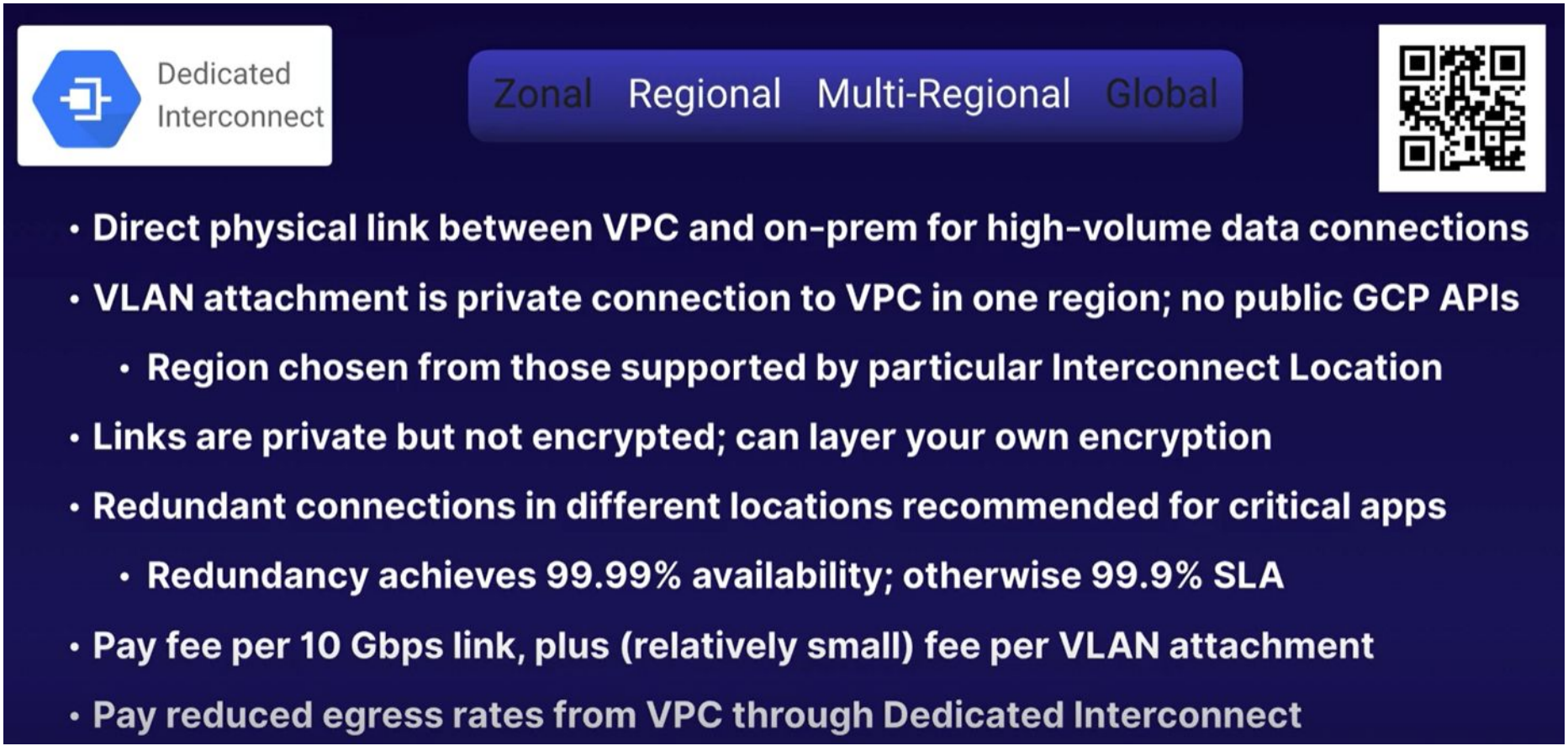


Internal Networking: Cloud VPN

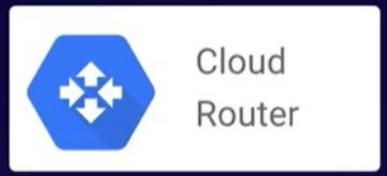
Zonal Regional Multi-Regional Global

- IPsec VPN to connect to VPC via public internet for low-volume data connections
- For persistent, static connections between gateways (i.e. not for a dynamic client)
  - Peer VPN gateway must have static (unchanging) IP
- Encrypted link to VPC (as opposed to Dedicated Interconnect), into one subnet
- Supports both static and dynamic routing
- 99.9% availability SLA
- Pay per tunnel-hour

Internal Networking: Dedicated Interconnect



Internal Networking: Cloud Router



Zonal Regional Multi-Regional Global

- Dynamic routing (BGP) for hybrid networks linking GCP VPCs to external networks
- Works with Cloud VPN and Dedicated Interconnect
- Automatically learns subnets in VPC and announces them to on-prem network
- Without Cloud Router you must manage static routes for VPN
  - Changing the IP addresses on either side of VPN requires recreating it
- Free to set up
- Pay for usual VPC egress

Internal Networking: CDN Interconnect

Zonal Regional Multi-Regional Global

- Direct, low-latency connectivity to certain CDN providers, with cheaper egress
- For external CDNs, not Google's Cloud CDN service
  - Supports Akamai, Cloudflare, Fastly, and more
- Works for both pull and push cache fills
  - Because it's for all traffic with that CDN
- Contact CDN provider to set up for GCP project and which regions
- Free to enable, then pay less for the egress you configured

Hands-on lab

Design for security

Building and Managing

Identity and Access - Core Security

Roles

Zonal Regional Multi-Regional Global

- Roles are collections of Permissions to use or manage GCP resources

AWS IAM Policies

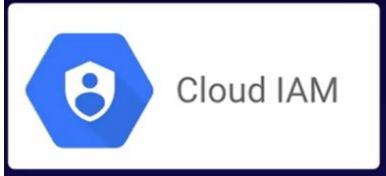
Identity and Access - Roles

Zonal Regional Multi-Regional Global

- Roles are collections of Permissions to use or manage GCP resources
- Permissions allow you to perform certain actions: `Service.Resource.Verb`
- Primitive Roles: Owner, Editor, Viewer
  - Viewer is read-only; Editor can change things; Owner can control access & billing
  - Pre-date IAM service, may still be useful (e.g. dev/test envs), but often too broad
- Predefined Roles: Give granular access to specific GCP resources (IAM)
  - E.g.: `roles/bigquery.dataEditor`, `roles/pubsub.subscriber`
- Custom Roles: Project- or Org-level collections you define of granular permissions

Identity and Access - Cloud IAM

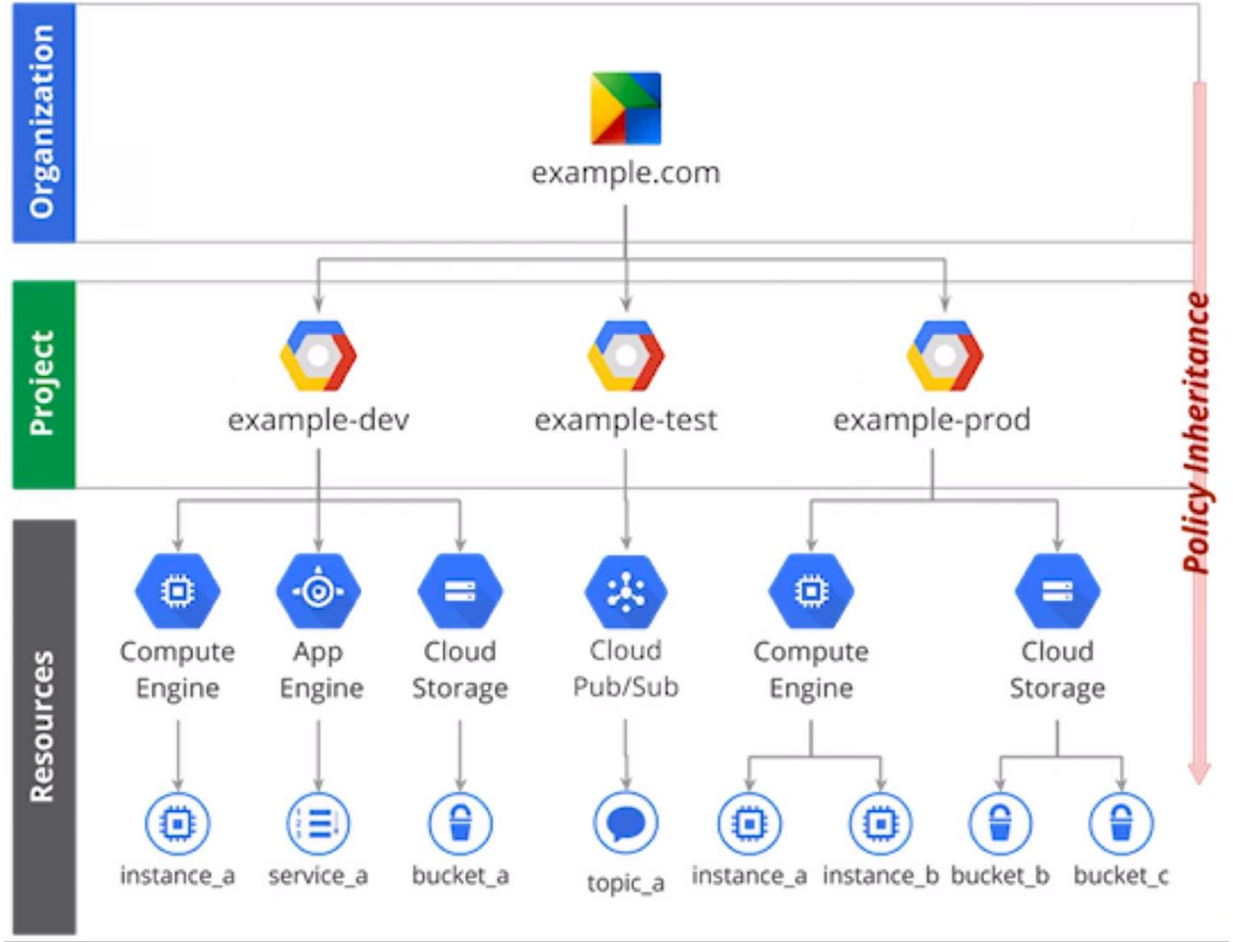
Identity and Access - Cloud IAM



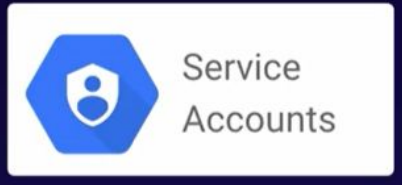
Zonal Regional Multi-Regional Global

- Controls access to GCP resources: authorization, not really authentication/identity
- Member is user, group, domain, service account, or the public (e.g. “allUsers”)
  - Individual Google account, Google group (👍), G Suite / Cloud Identity domain
  - Service account (👍) belongs to application/instance, not individual end user
  - Every identity has a unique e-mail address, including service accounts
- Policies bind Members to Roles at a hierarchy level: Org, Folder, Project, Resource
  - Answer: *Who* can do *what* to *which thing(s)*?
- IAM is free; pay for authorized GCP service usage

Identity and Access - Core Security



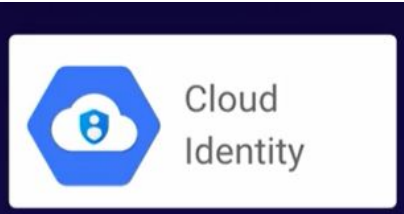
Identity and Access - Service Accounts



Zonal Regional Multi-Regional Global

- Special type of Google account that represents an application, not an end user
- Can be “assumed” by applications or individual users (when so authorized)
- “Important: For almost all cases, whether you are developing locally or in a production application, you should use service accounts, rather than user accounts or API keys.”
- Consider resources and permissions required by application; use least privilege
- Can generate and download private keys (user-managed keys), for non-GCP, but...
- Cloud-Platform-managed keys (👍) are better, for GCP (i.e. GCF, GAE, GCE, and GKE)
  - No direct downloading: Google manages private keys & rotates them once a day

Identity and Access - Cloud Identity



Zonal Regional Multi-Regional Global

- Identity as a Service (IDaaS, not DaaS) to provision and manage users and groups

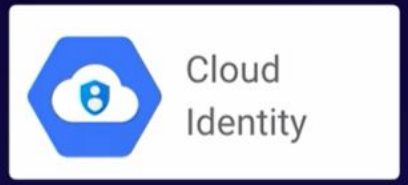
AWS IAM

G Suite

GMail/Google Account

Active Directory

Identity and Access - Core Security



Zonal Regional Multi-Regional Global

- Identity as a Service (IDaaS, not DaaS) to provision and manage users and groups
- Free Google Accounts for non-G-Suite users, tied to a verified domain
- Centrally manage all users in Google Admin console; supports compliance
- 2-Step verification (2SV/MFA) and enforcement (👍), including security keys
- Sync from Active Directory and LDAP directories via Google Cloud Directory Sync
- Identities work with other Google services (e.g. Chrome)
- Identities can be used to SSO with other apps via OIDC, SAML, OAuth2
- Cloud Identity is free; pay for authorized GCP service usage

Identity and Access - Security Key Enforcement

Zonal Regional Multi-Regional Global

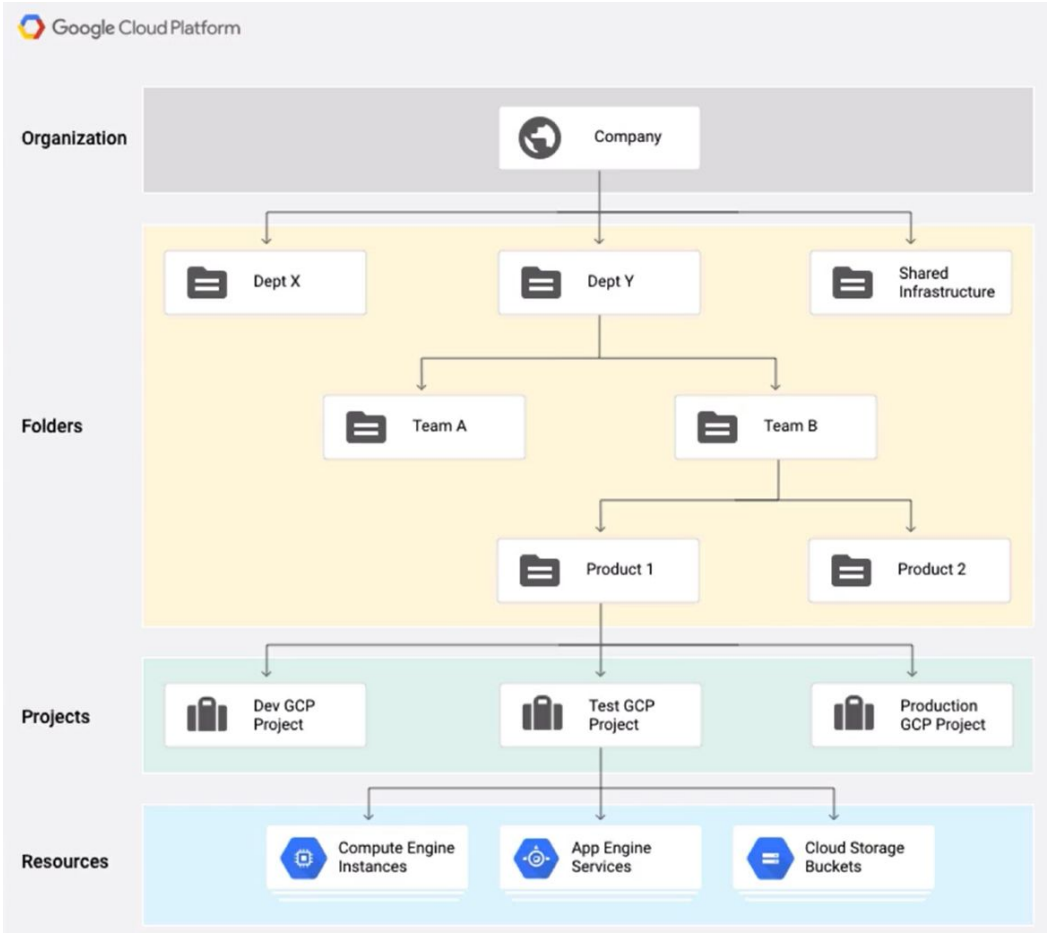
- USB or Bluetooth 2-step verification device that prevents phishing
- Not like just getting a code via email or text message...
- Device also verifies the target service
- Eliminates man-in-the-middle (MITM) attacks against GCP credentials

Identity and Access - Resource Manager

Zonal Regional Multi-Regional Global

- Centrally manage & secure organization's projects with custom folder hierarchy
- Organization resource is root node in hierarchy; folders per your business needs
- Tied 1:1 to a Cloud Identity / G Suite domain, then owns all newly-created projects
  - Without this organization, specific identities (people) must own GCP projects
- "Recycle bin" allows undeleting projects
- Define custom IAM roles at org level
- Apply IAM policies at organization, folder, or project levels

Identity and Access - Resource Manager

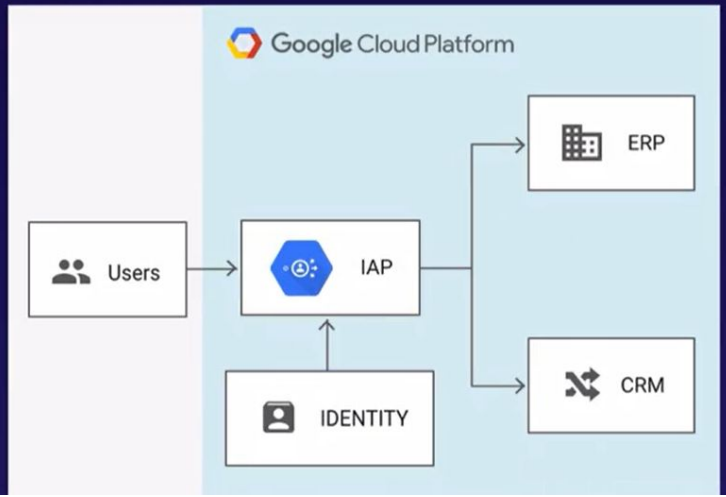


Identity and Access - Cloud IAP

Cloud IAP

Zonal Regional Multi-Regional Global

- Guards apps running on GCP via identity verification, not VPN access
- Based on CLB & IAM, and only passes authed requests through



Identity and Access - Cloud IAP

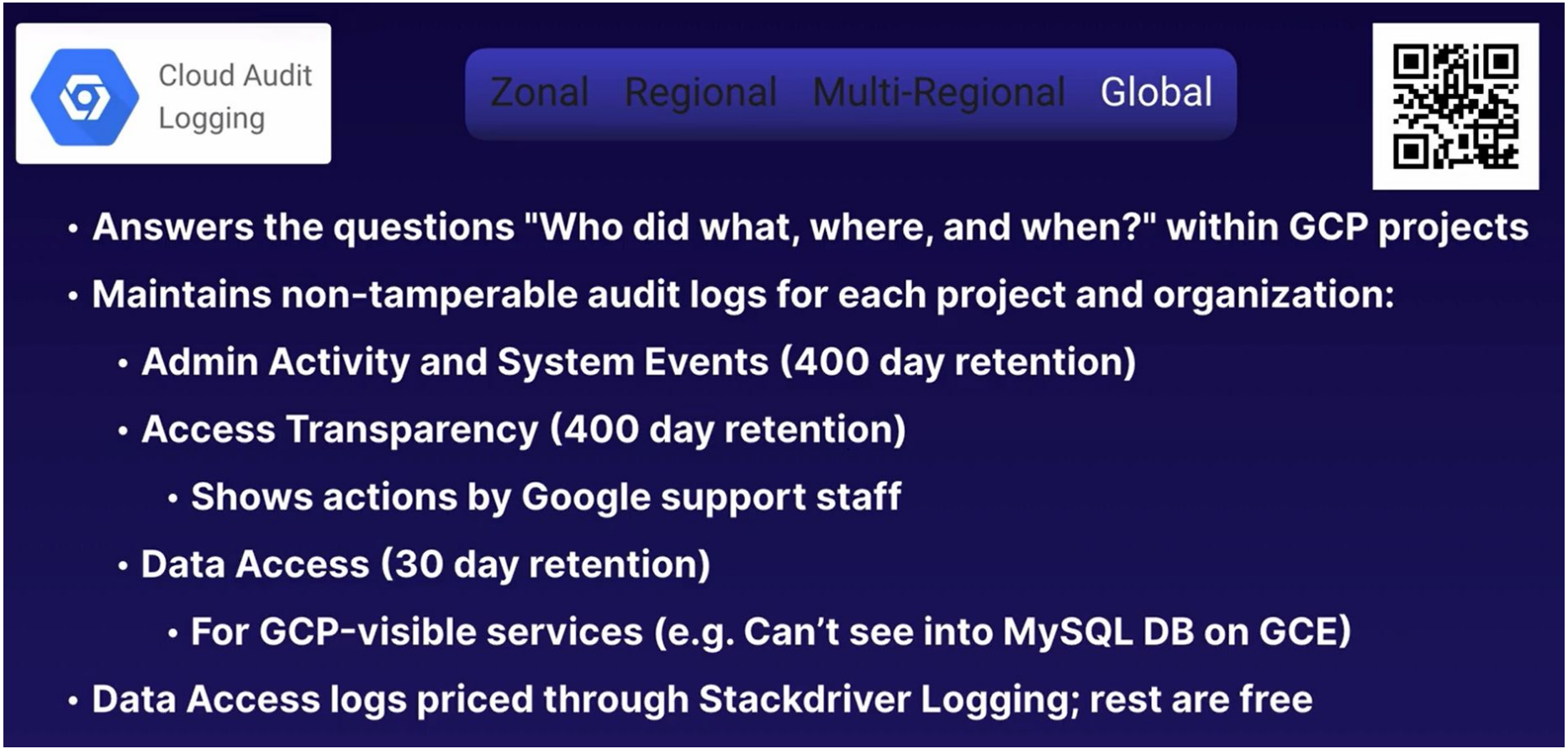
Cloud IAP

Zonal Regional Multi-Regional Global

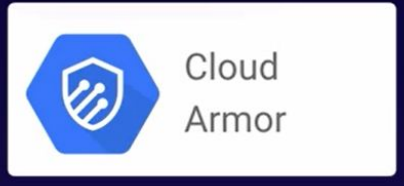
- Guards apps running on GCP via identity verification, not VPN access
- Based on CLB & IAM, and only passes authed requests through

Amazon API Gateway

Identity and Access - Cloud Audit Logging



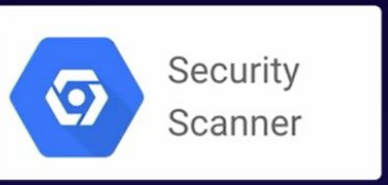
Security Management : Cloud Armor



Zonal Regional Multi-Regional Global

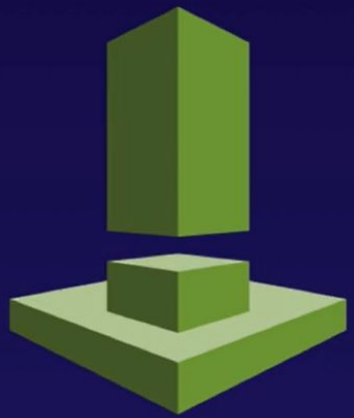
- Edge-level protection from DDoS & other attacks on global HTTP(S) LB
- Offload work: Blocked attacks never reach your systems
- Monitor: Detailed request-level logs available in Stackdriver Logging
- Manage IPs with CIDR-based allow/block lists (aka whitelist/blacklist)
- More intelligent rules forthcoming (e.g. XSS, SQLi, geo-based, custom)
- Preview effect of changes before making them live
- Pay per policy and rule configured, plus for incoming request volume

Security Management: Security Scanner



Zonal Regional Multi-Regional Global

- Free but limited GAE app vulnerability scanner with “very low false positive rates”

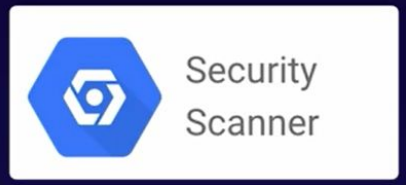


Amazon Inspector

Trustwave App Scanner

Qualys Web Application Scanning

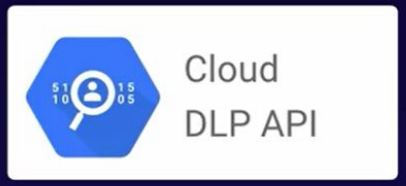
Security Management: Security Scanner



Zonal Regional Multi-Regional Global

- Free but limited GAE app vulnerability scanner with “very low false positive rates”
- “After you set up a scan, Cloud Security Scanner automatically crawls your application, following all links within the scope of your starting URLs, and attempts to exercise as many user inputs and event handlers as possible.”
- Can identify:
  - Cross-site-scripting (XSS)
  - Flash injection
  - Mixed content (HTTP in HTTPS)
  - Outdated/insecure libraries

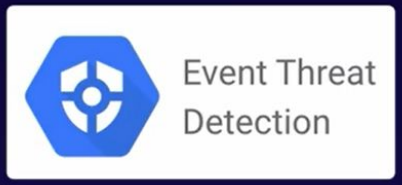
Security Management: Cloud DLP API



Zonal Regional Multi-Regional Global

- Finds and optionally redacts sensitive info in unstructured data streams
- Helps you minimize what you collect, expose, or copy to other systems
- 50+ sensitive data detectors, including: credit card numbers, names, social security numbers, passport numbers, driver's license numbers (US and some other jurisdictions), phone numbers, and other personally identifiable information (PII)
- Data can be sent directly, or API can be pointed at GCS, BQ, or Cloud DataStore
- Can scan both text and images
- Pay for amount of data processed (per GB)—and gets cheaper when large volume
  - Pricing for storage now very simple (June 2019), but for streaming is still a mess

Security Management: Event Threat Detection

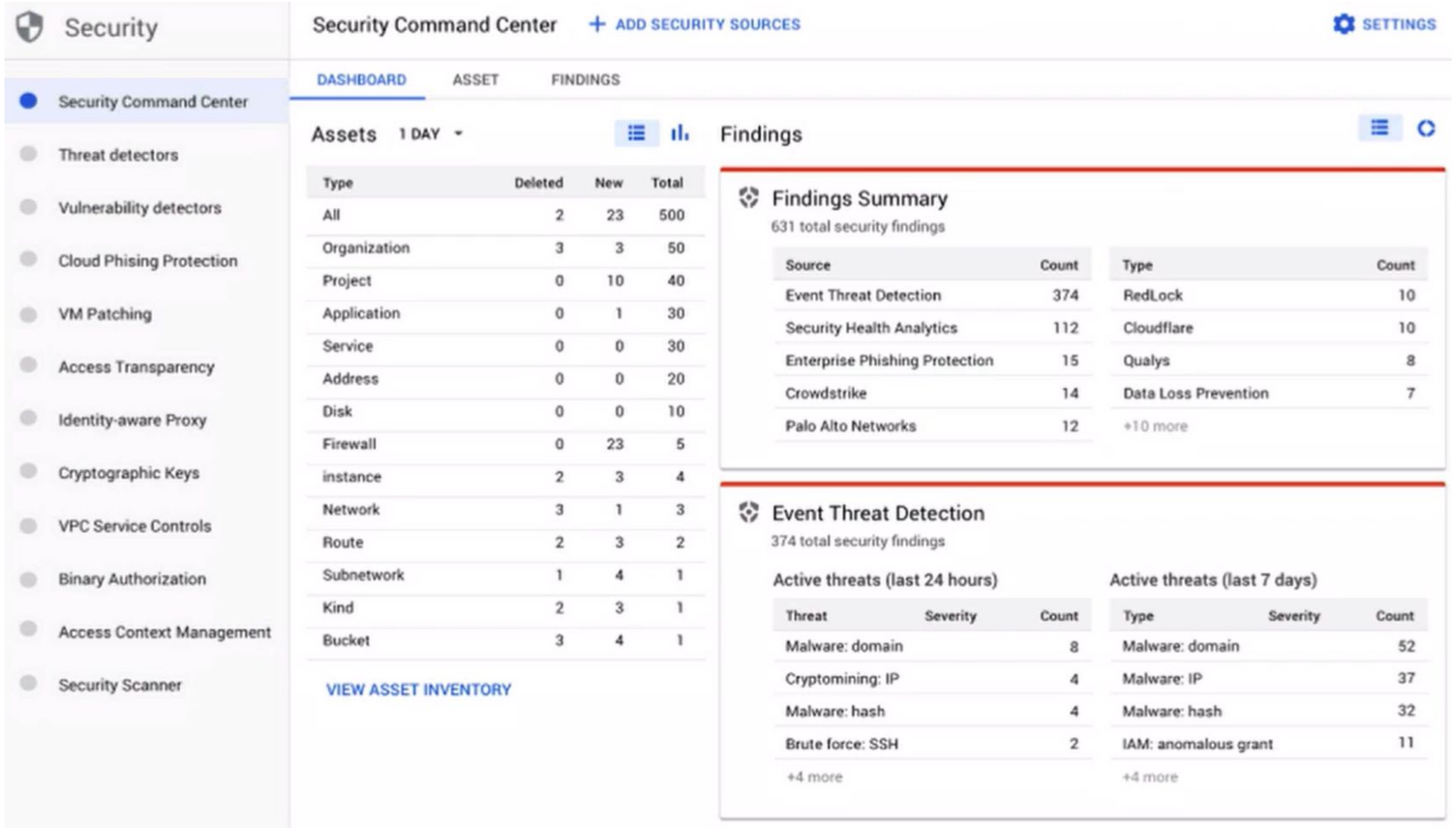


Zonal Regional Multi-Regional Global

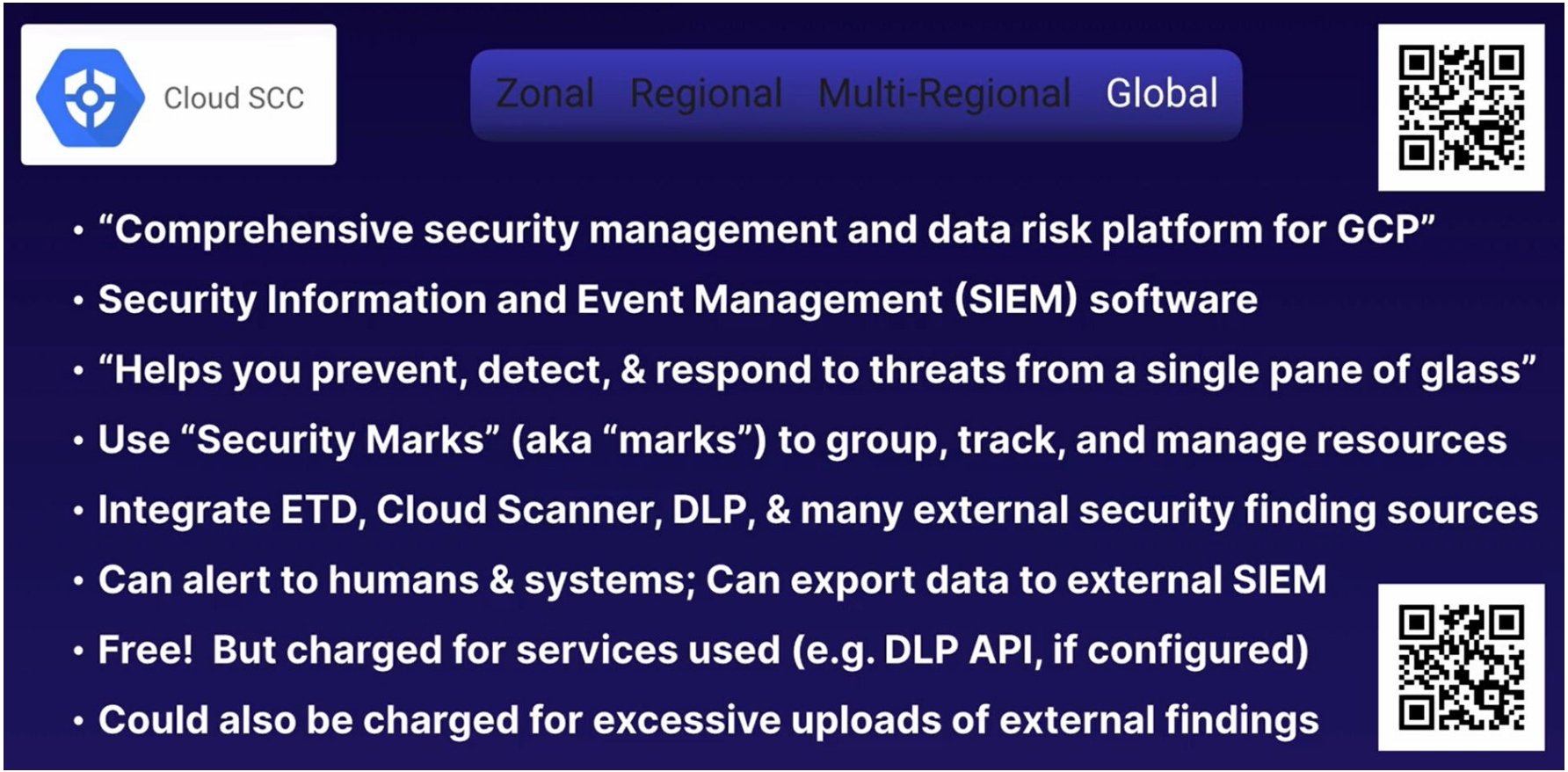
- Automatically scans your Stackdriver logs for suspicious activity
- Uses industry-leading threat intelligence, including Google Safe Browsing
- Quickly detects many possible threats, including:
  - Malware, cryptomining, outgoing DDoS attacks, port scanning, brute-force SSH
  - Also: Unauthorized access to GCP resources via abusive IAM access
- Can export parsed logs to BigQuery for forensic analysis
- Integrates with SIEMs like Google's Cloud SCC or via Cloud Pub/Sub
- No charge for ETD, but charged for its usage of other GCP services (like SD Logging)

Security Management: Cloud SCC

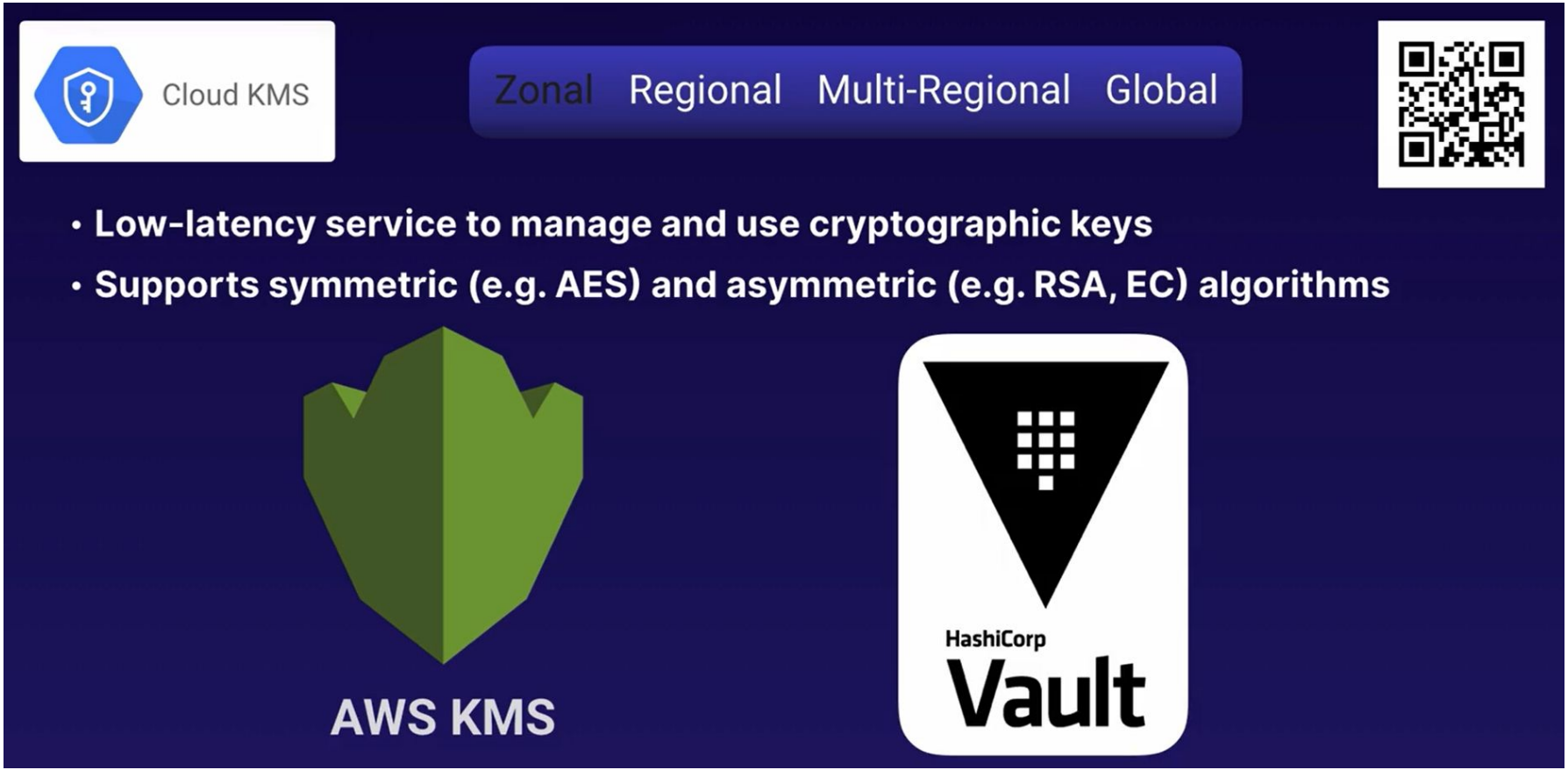
Security Management: Cloud SCC



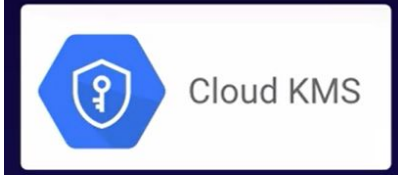
Security Management: Cloud SCC



Encryption Key Management: Cloud KMS



Encryption Key Management: Cloud KMS



Zonal Regional Multi-Regional Global

- Low-latency service to manage and use cryptographic keys
- Supports symmetric (e.g. AES) and asymmetric (e.g. RSA, EC) algorithms
- Move secrets out of code (and the like) and into the environment, in a secure way
- Integrated with IAM & Cloud Audit Logging to authorize & track key usage
- Rotate keys used for new encryption either automatically or on demand
  - Still keeps old active key versions, to allow decrypting
- Key deletion has 24 hour delay, "to prevent accidental or malicious data loss"
- Pay for active key versions stored over time
- Pay for key use operations (i.e. encrypt/decrypt; admin operations are free)

Encryption Key Management: Cloud HSM

Cloud HSM

Zonal Regional Multi-Regional Global

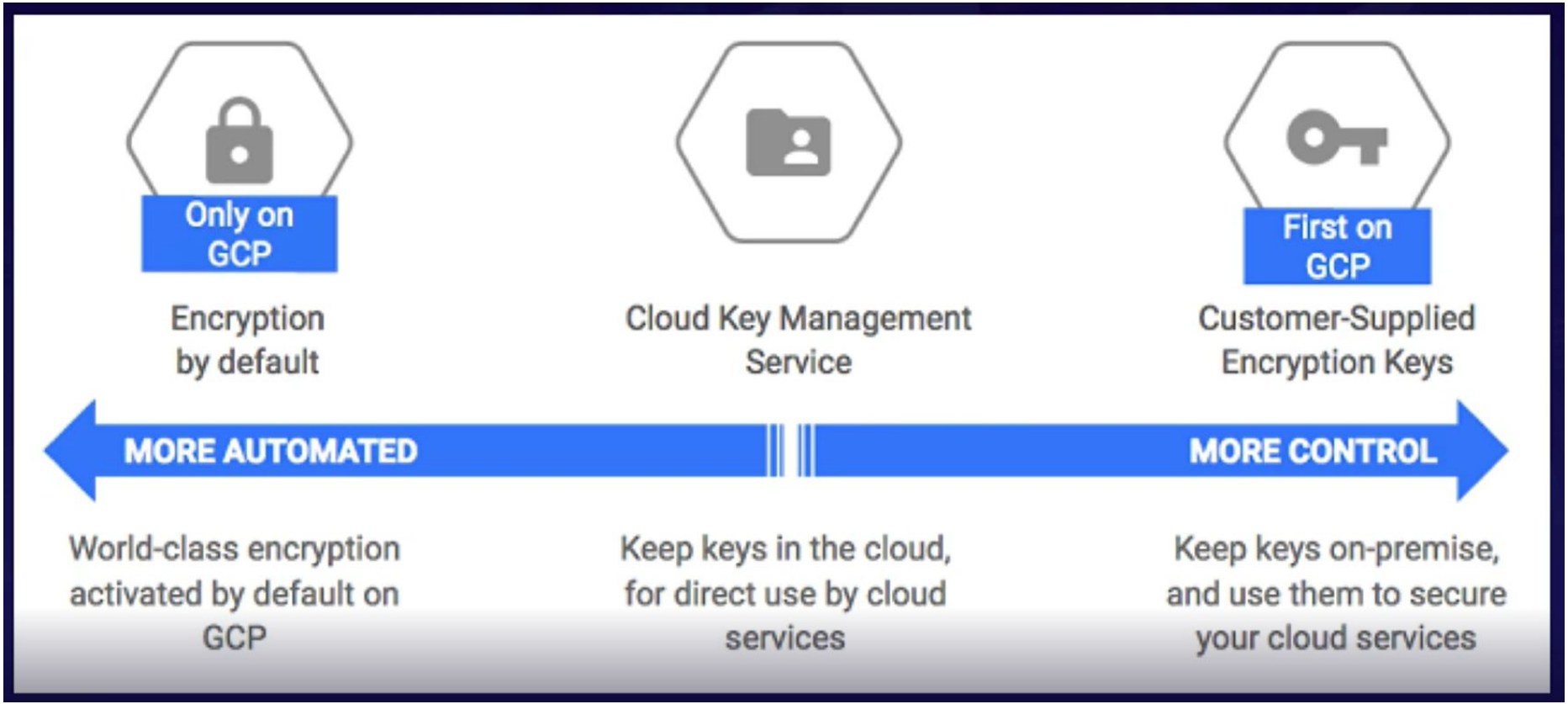
- Cloud KMS keys managed by FIPS 140-2 Level 3 certified HSMs
- Device hosts encryption keys and performs cryptographic operations
- Enables you to meet compliance that mandates hardware environment

AWS CloudHSM

SafeNet HSM

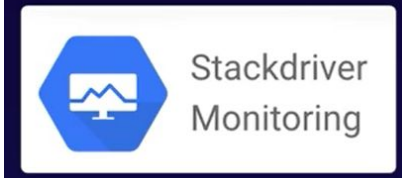
Encryption Key Management: Cloud KMS

Encryption Key Management Options: Flowchart



Encryption Key Management Options: Flowchart

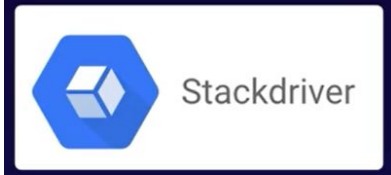
Operations and Management



Zonal Regional Multi-Regional Global

- Gives visibility into perf, uptime, & overall health of cloud apps (based on collectd)
- Includes built-in/custom metrics, dashboards, global uptime monitoring, & alerts
- Follow the trail: Links from alerts to dashboards/charts to logs to traces
- Cross-cloud: GCP, of course, but monitoring agent also supports AWS
- Alerting policy config includes multi-condition rules & resource organization
- Alert via email, GCP Mobile App, SMS, Slack, PagerDuty, AWS SNS, webhook, etc.
- Automatic GCP/Anthos metrics always free
- Pay for API calls & per MB for custom or AWS metrics

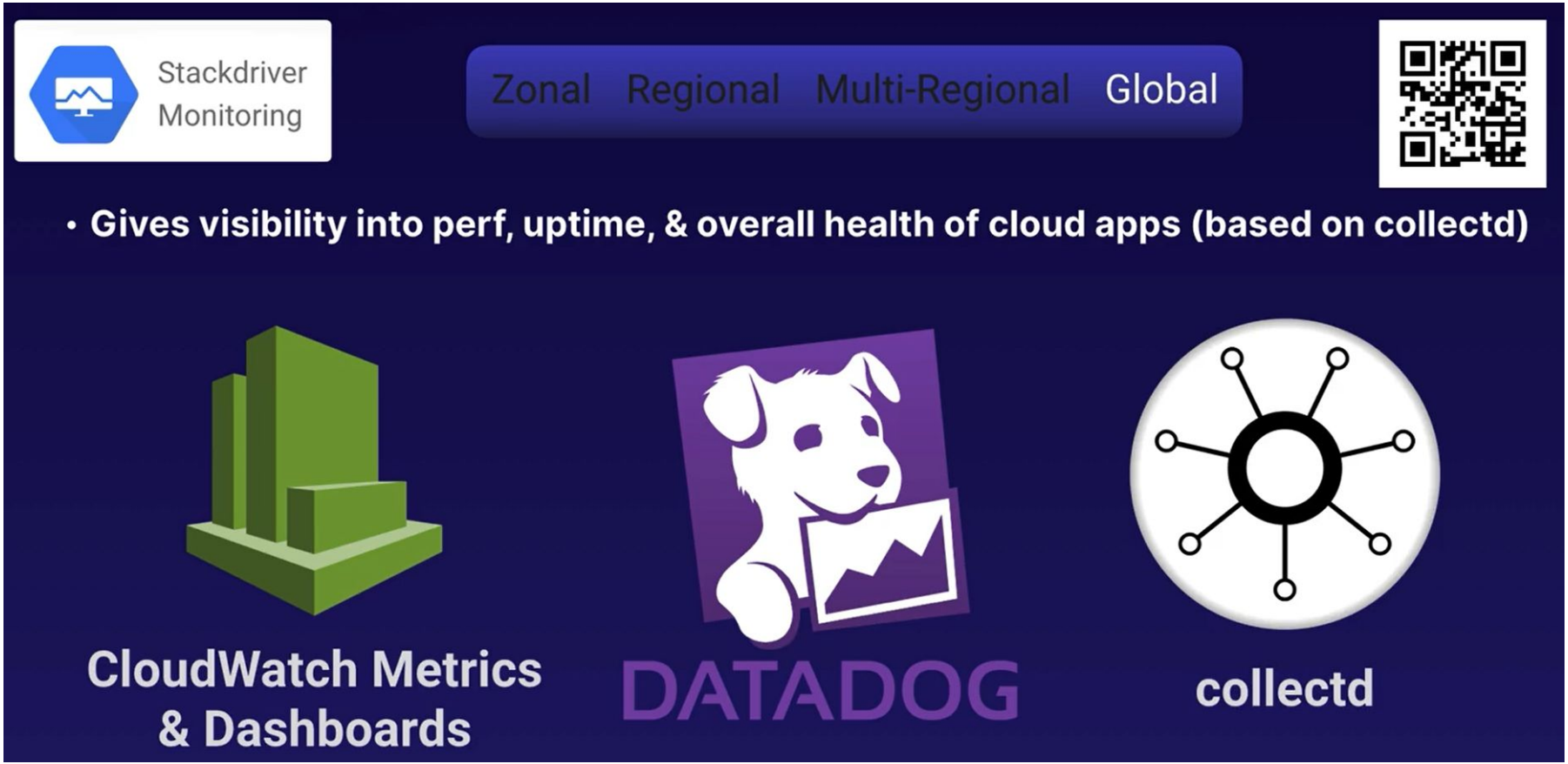
Operations and Management



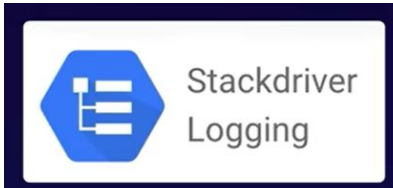
Zonal Regional Multi-Regional Global

- Family of services for monitoring, logging, & diagnosing apps on GCP/AWS/hybrid
- Service integrations add lots of value—among Stackdriver and with GCP
- One Stackdriver account can track multiple:
  - GCP projects
  - AWS accounts
  - Other resources
- Simple usage-based pricing
  - No longer previous system of tiers, allotments, and overages

Operations and Management



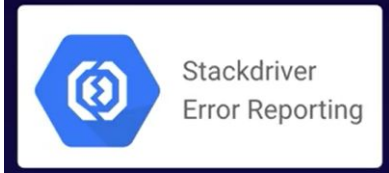
Operations and Management



Zonal Regional Multi-Regional Global

- Store, search, analyze, monitor, and alert on log data & events (based on Fluentd)
- Collection built into some GCP, AWS support with agent, or custom send via API
- Debug issues via integration with Stackdriver Monitoring, Trace & Error Reporting
- Create real-time metrics from log data, then alert or chart them on dashboards
- Send real-time log data to BigQuery for advanced analytics and SQL-like querying
- Powerful interface to browse, search, and slice log data
- Export log data to GCS to cost-effectively store log archives
- Pay per GB ingested & stored for one month, but first 50GB/project free
- Cloud Audit Logging / Access Transparency logs also free

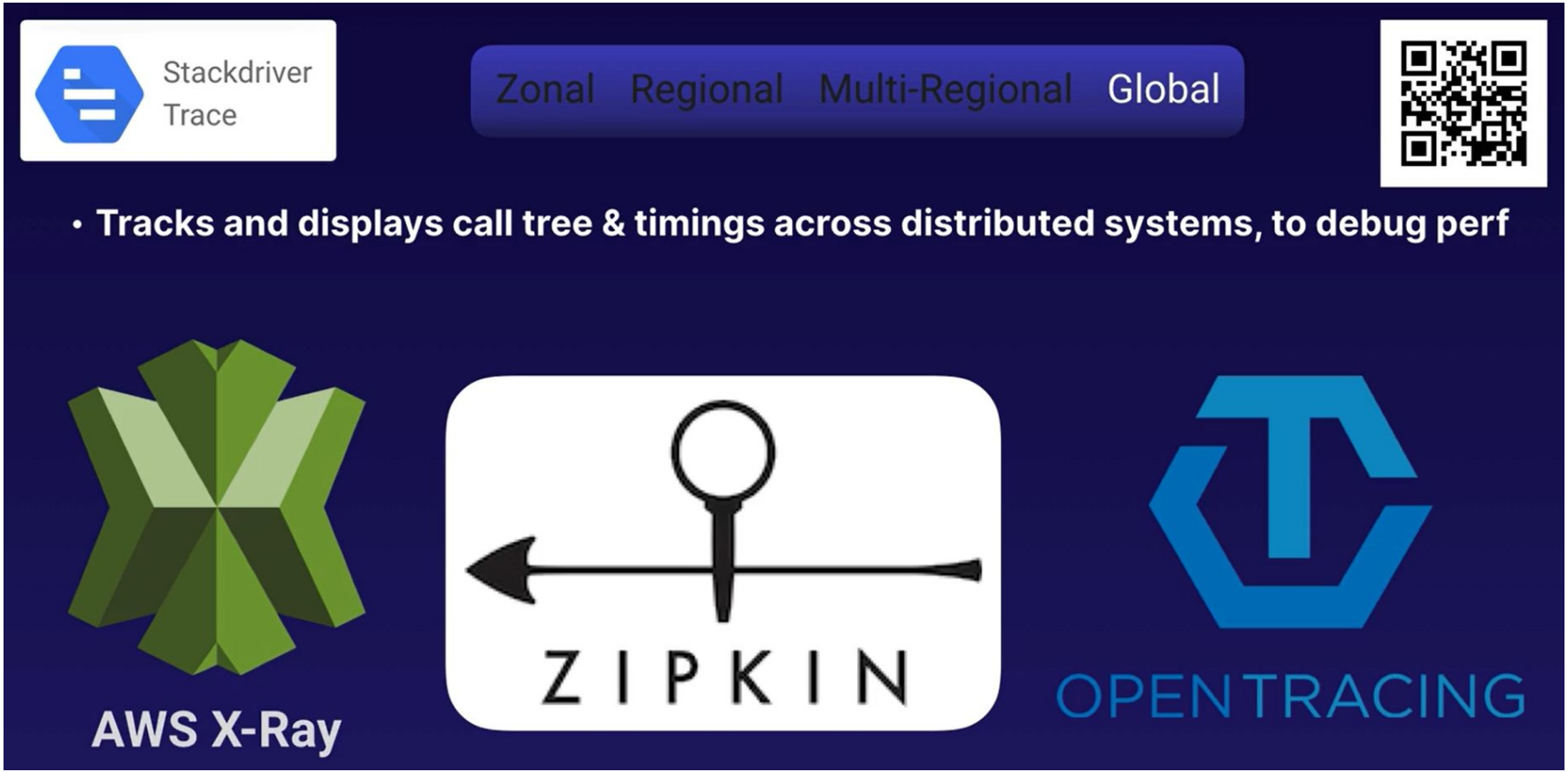
Operations and Management



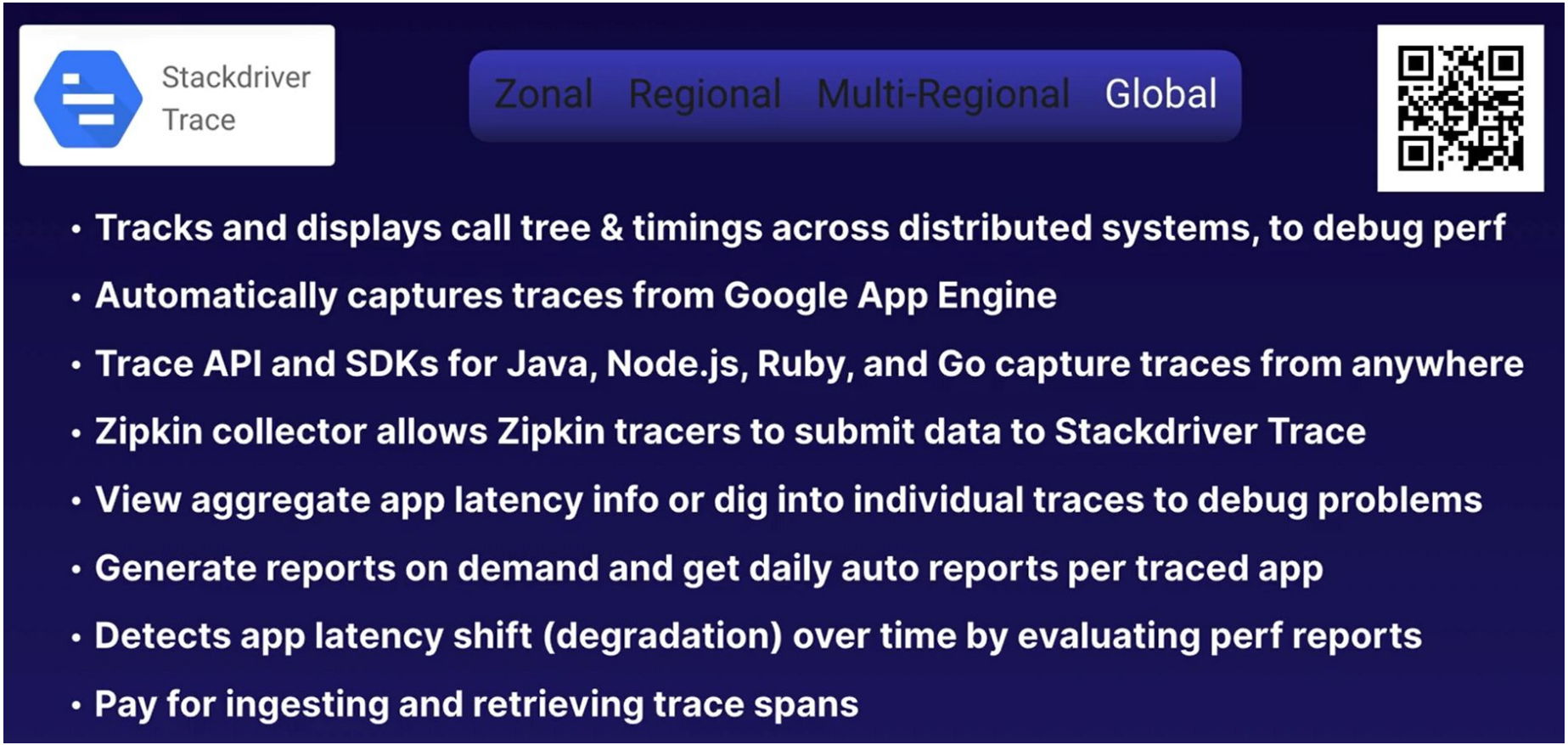
Zonal Regional Multi-Regional Global

- Counts, analyzes, aggregates, & tracks crashes in helpful centralized interface
- Smartly aggregates errors into meaningful groups tailored to language/framework
- Instantly alerts when a new app error cannot be grouped with existing ones
- Link directly from notifications to error details:
  - Time chart, occurrences, affected user count, first/last seen dates, cleaned stack
- Exception stack trace parser knows Java, Python, JavaScript, Ruby, C#, PHP, & Go
- Jump from stack frames to source to start debugging
- No direct charge; pay for source data in Stackdriver Logging

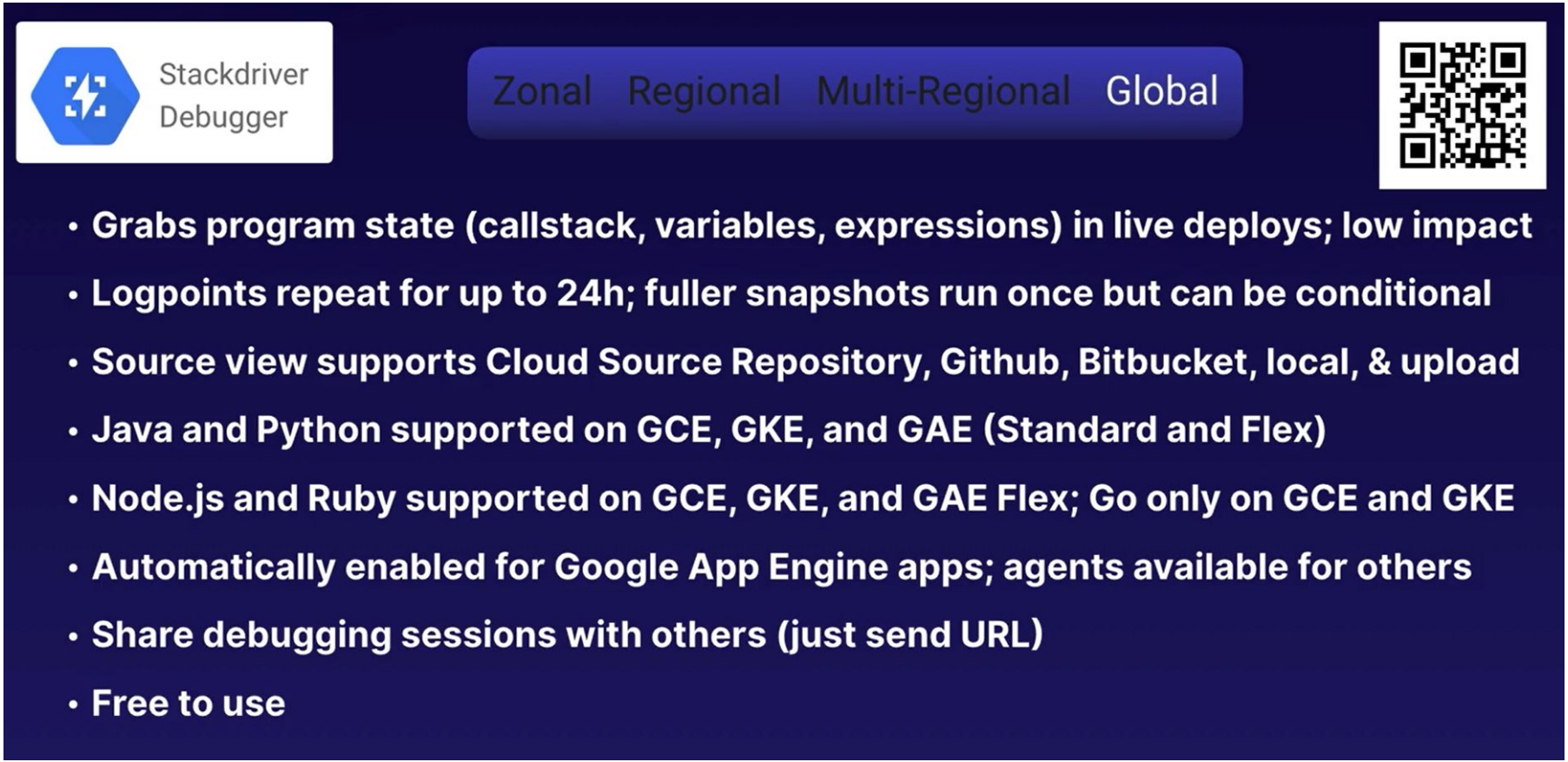
Operations and Management



Operations and Management



Operations and Management

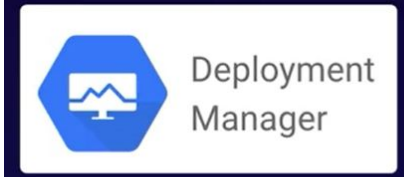


Operations and Management

Zonal Regional Multi-Regional Global

- Continuous CPU and memory profiling to improve perf & reduce cost
- Low overhead (Typical: 0.5%; Max: 5%)—so use it in prod, too!
- Supports Go, Java, Node.js, and Python (3.2+)
- Agent-based
- Saves profiles for 30 days
- Can download profiles for longer-term storage
- Free to use

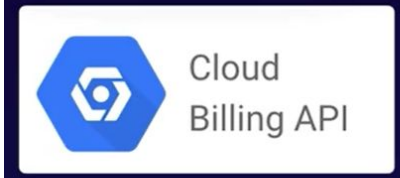
Operations and Management



Zonal Regional Multi-Regional Global

- Create/manage resources via declarative templates: “Infrastructure as Code”
- Declarative allows automatic parallelization
- Templates written in YAML, Python, or Jinja2
- Supports input and output parameters, with JSON schema
- Create and update of deployments both support preview
- Free service; just pay for resources involved in deployments

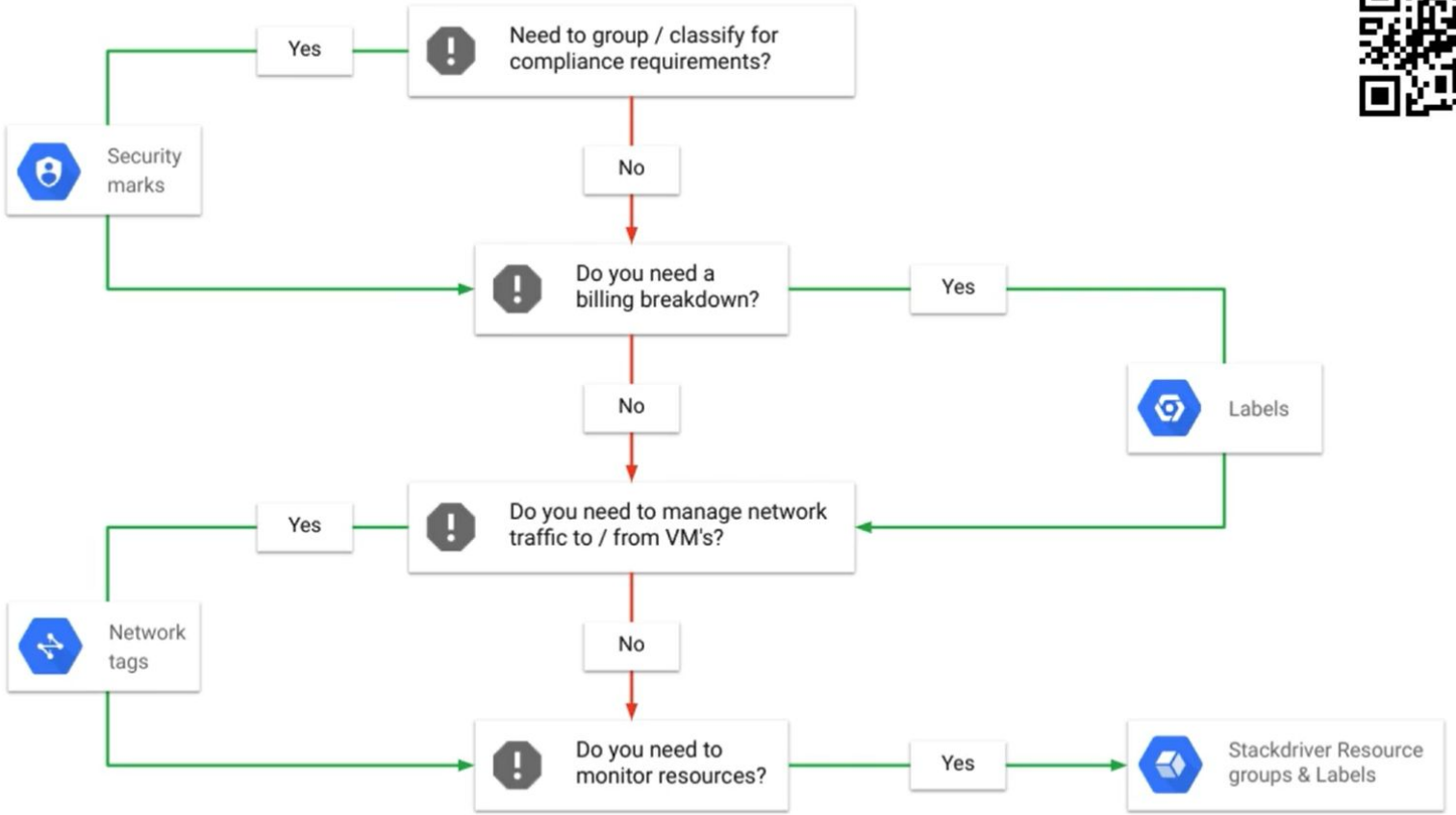
Operations and Management



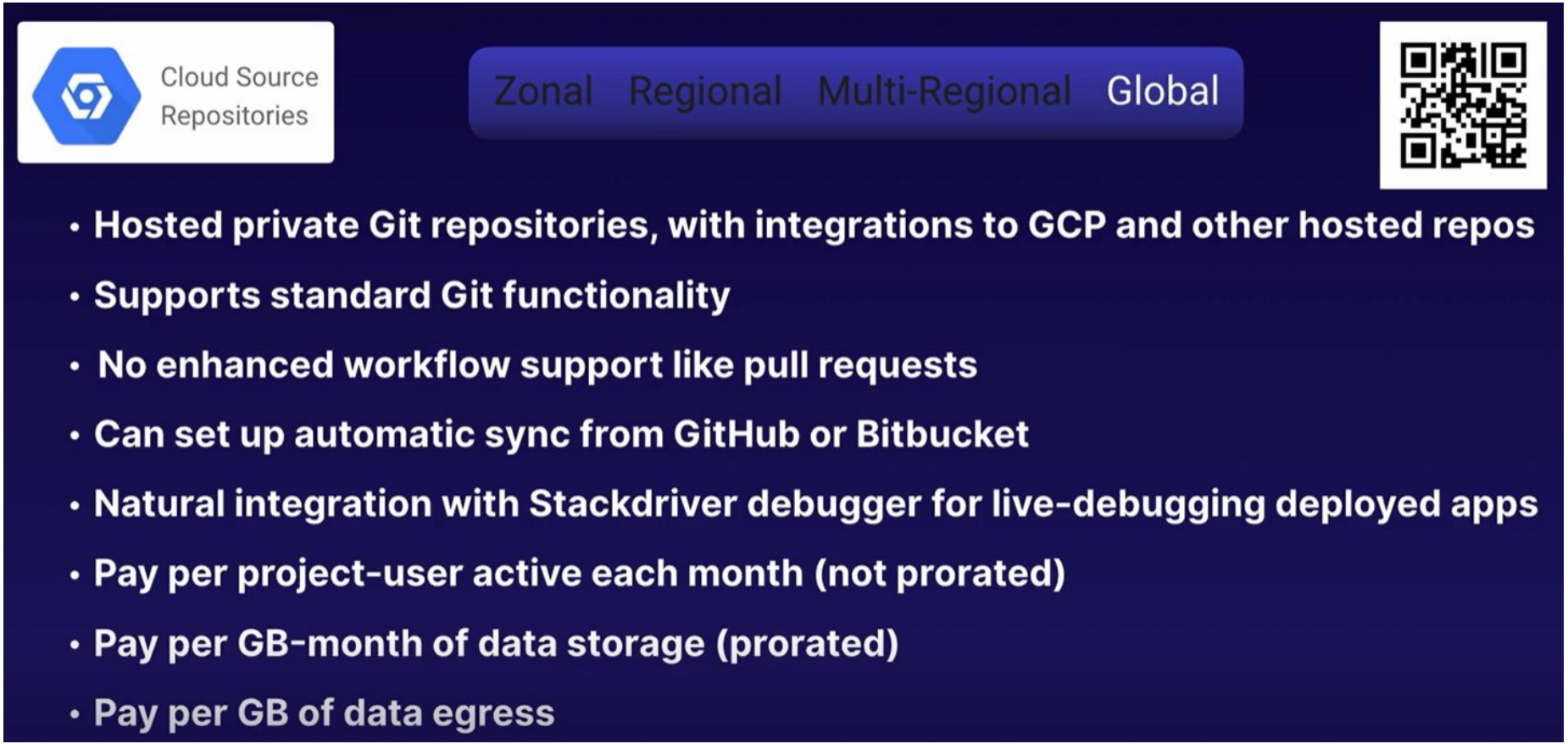
Zonal Regional Multi-Regional Global

- Programmatically manage billing for GCP projects and get GCP pricing
- Billing config
  - List billing accounts; get details and associated projects for each
  - Enable (associate), disable (disassociate), or change project's billing account
- Pricing
  - List billable SKUs; get public pricing (including tiers) for each
  - Get SKU metadata like regional availability
- Export of current bill to GCS or BQ is possible—but configured via console, not API

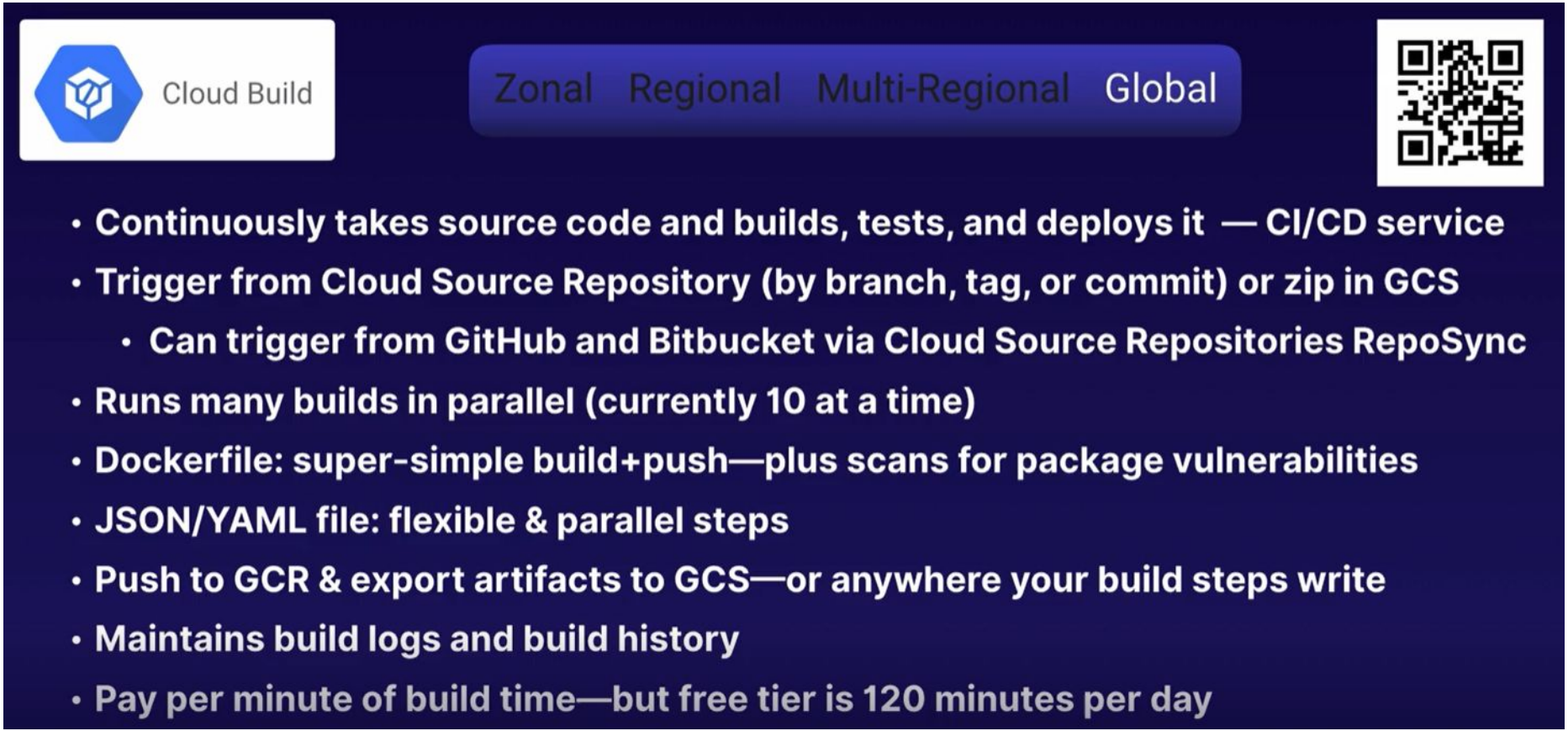
Resource Annotation Options: Flowchart



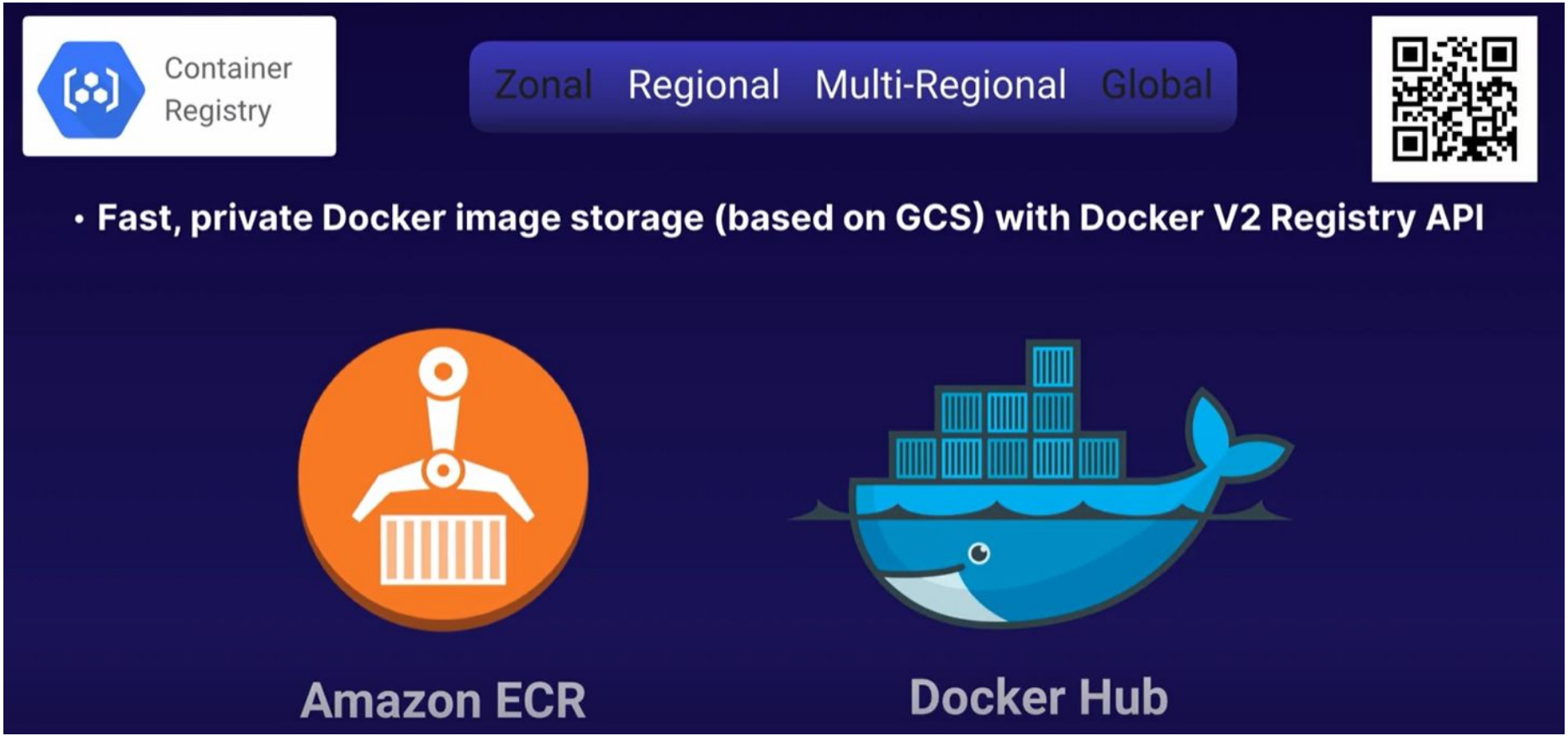
Development and APIs



Development and APIs



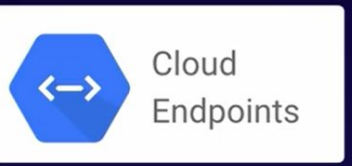
Development and APIs



Development and APIs

Development and APIs

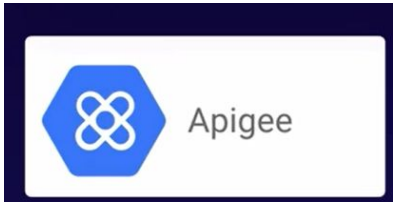
Development and APIs



Zonal Regional Multi-Regional Global

- Handles authorization, monitoring, logging, & API keys for APIs backed by GCP
- Proxy instances are distributed and hook into Cloud Load Balancer
- Super-fast Extensible Service Proxy (ESP) container based on nginx: <1 ms / call
- Uses JWTs and integrates with Firebase, Auth0, & Google Auth
- Integrates with Stackdriver Logging and Stackdriver Trace
- Extensible Service Proxy (ESP) can transcode HTTP/JSON to gRPC
  - But API needs to be resource-oriented (i.e. RESTful)
- Pay per call to your API

Development and APIs



Zonal Regional Multi-Regional Global

- Full-featured & enterprise-scale API management platform for whole API lifecycle
- Transform calls between different protocols: SOAP, REST, XML, binary, custom
- Authenticate via OAuth/SAML/LDAP; authorize via Role-Based Access Control
- Throttle traffic with quotas, manage API versions, etc.
- Apigee Sense identifies and alerts administrators to suspicious API behaviors
- Apigee API Monetization supports various revenue models / rate pans
- Team and Business tiers are flat monthly rate with API call quotas & feature sets
- “Enterprise” tier and special feature pricing are “Contact Sales”

Hands-on lab

Thank you