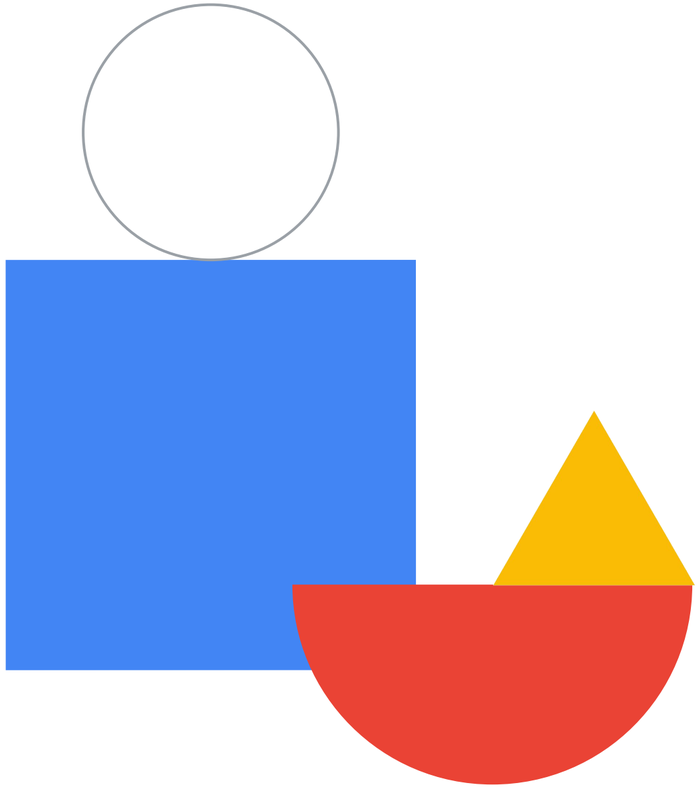
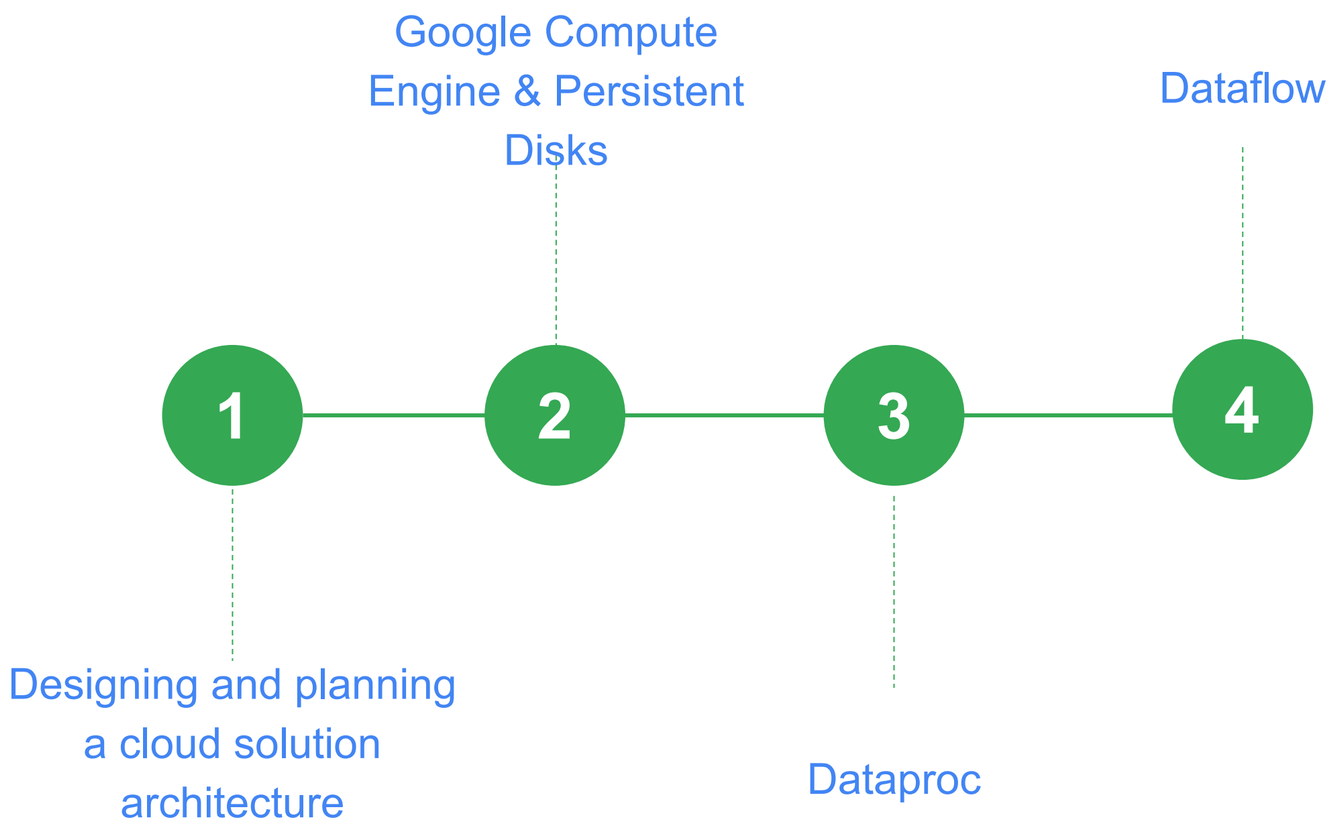
Google Cloud



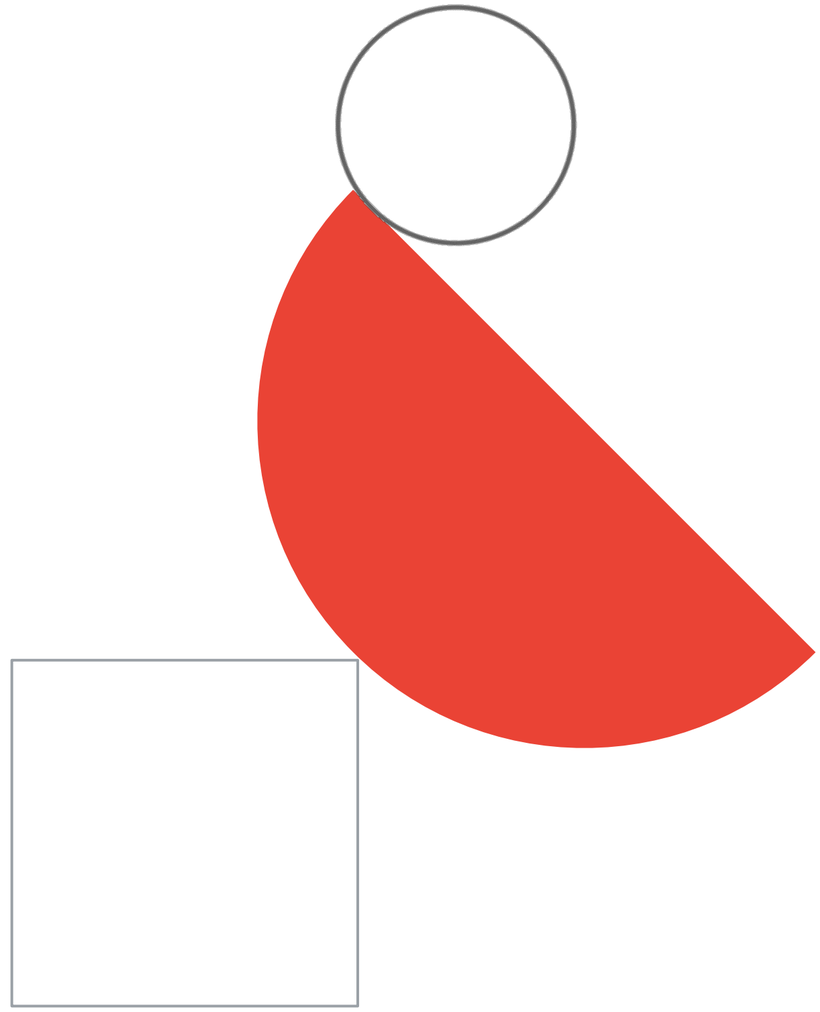
Preparing for Your Professional Cloud Architect Journey

Module 1: Designing and Planning a Cloud Solution Architecture

Week 2 topics



Designing and planning a cloud solution architecture



Define systems in scope for a cloud migration

… and / or decide on “cloud first” approach

Delivery by Drone

- Their website frontend, pilot, and truck management systems run on Kubernetes.
- Positional data for drone and truck location is kept in a MongoDB database clusters
- Drones stream video to virtual machines via stateful connection

Purchase & Product APIs

- APIs are simply built into monolithic apps, and were not designed for partner integration.
- APIs are running on Ubuntu linux VMs

Social Media Highlighting

- Single SuSE linux VM
- MySQL DB
- Redis
- Python

Define business and technical requirements

Business requirements

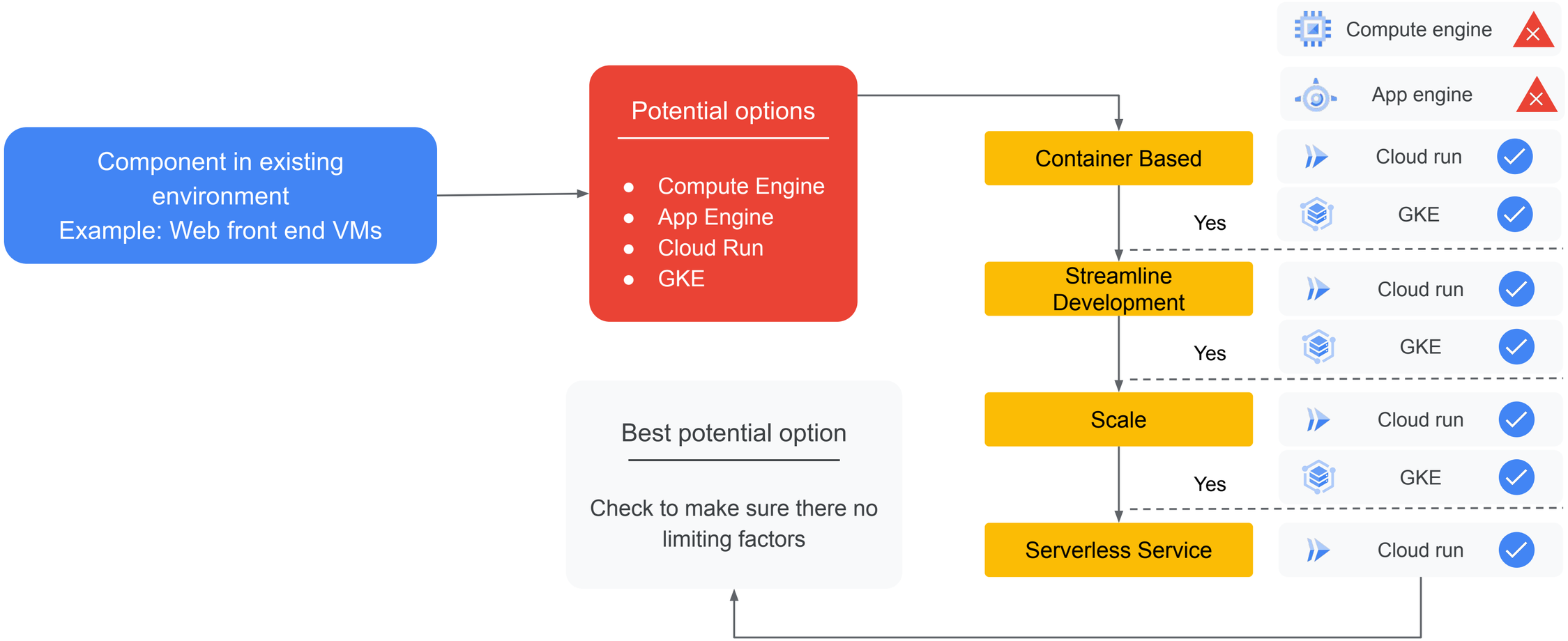
- Easily scale to handle additional demand when needed and expand to more test markets.
- Streamline development for application modernization and new features/products
- Ensure that developers spend as much time on core business functionality as possible, and not have to worry about scalability wherever possible
- Let partners order directly via API
- Deploy a production version of the social media highlighting service and ensure no inappropriate content

Technical requirements

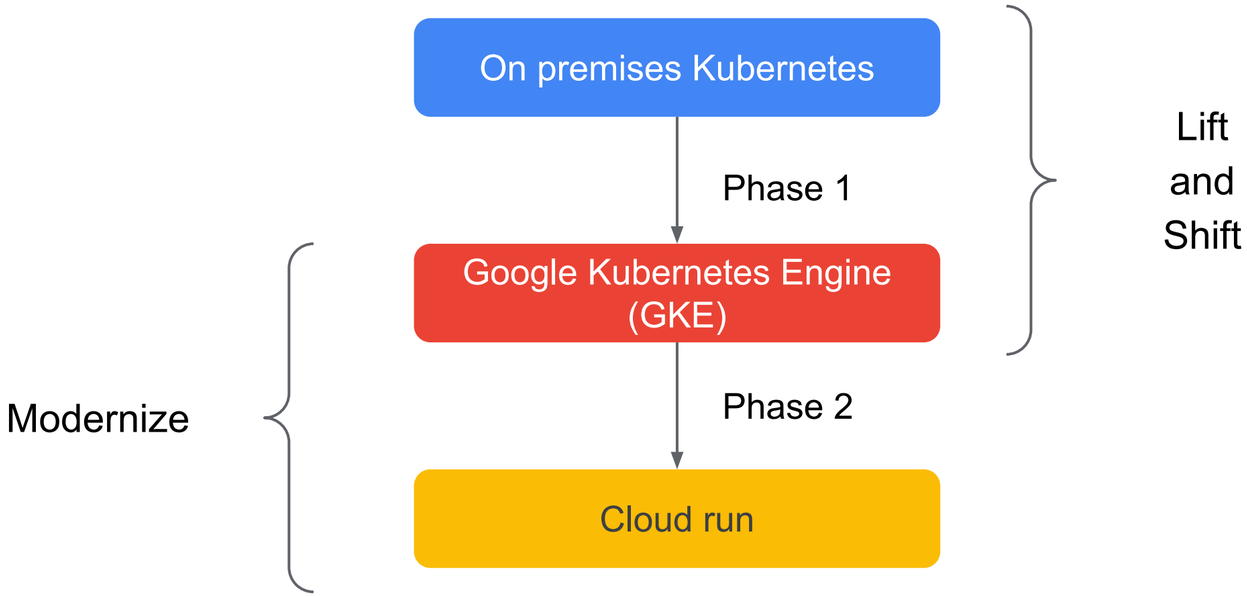
- Move to serverless services wherever possible
- Ensure that developers can deploy container-based workloads to testing and production environments in a highly scalable environment.
- Standardize on containers where possible, but also allow for existing virtualization infrastructure to run as-is without a re-write, so it can be slowly refactored over time
- Securely allow partner integration
- Stream IoT data from drones

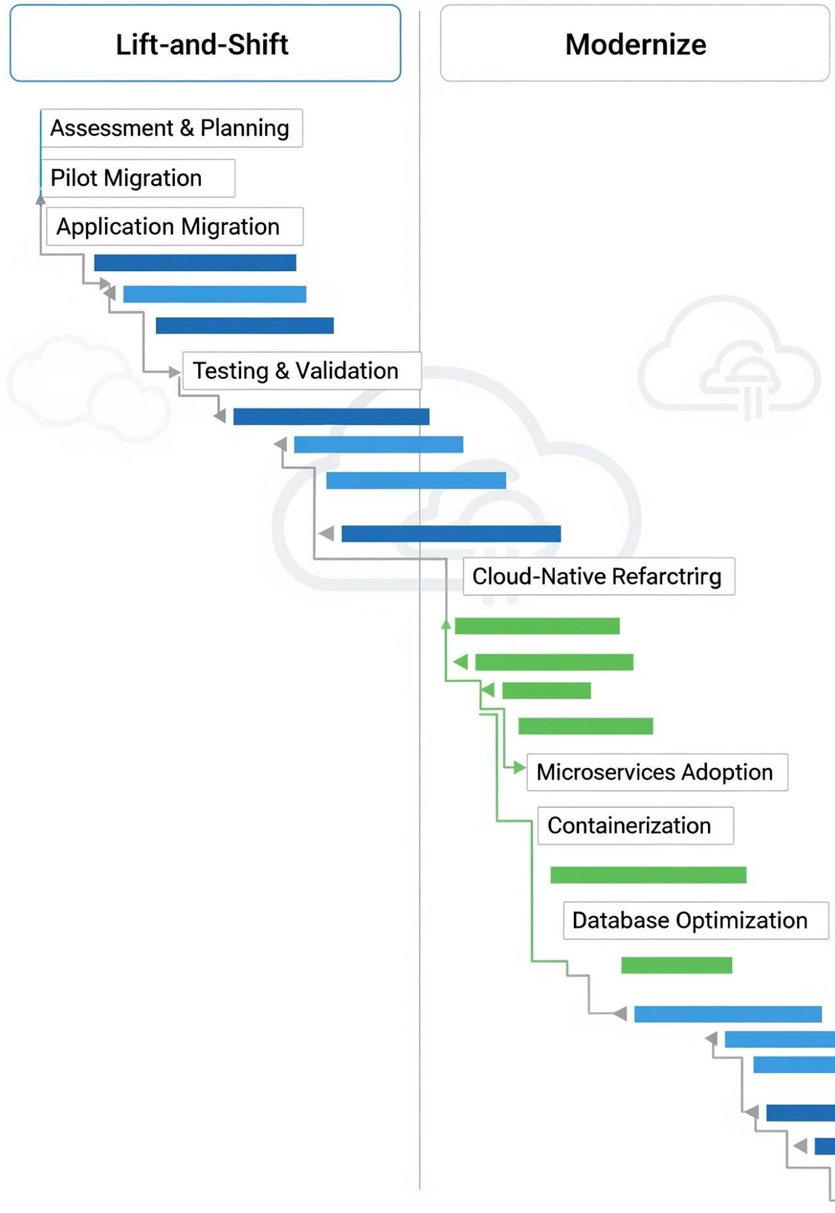
Mapping requirements to GCP resources

… through the lens of business & technical requirements



Planning for migration and the future





Migration guide & best practices

- Types of migrations and their use-cases
  - For example, “If the current app isn't meeting your goals—for example, you don't want to maintain it, it's too costly to migrate using one of the previously mentioned approaches, or it's not supported on Google Cloud—you can do a rebuild migration.”
- Building inventory of workloads in scope
  - … along with their dependencies!
- Best practices for validating a migration plan.

Diagnostic Question Discussion

You work for a large financial institution that is planning a multi-phase migration of its on-premises workloads to Google Cloud. The migration involves sensitive customer data and requires high availability and disaster recovery capabilities. You want to design a cloud solution architecture that meets these requirements while minimizing costs and ensuring compliance with industry regulations.

What should you do?

A. Migrate all workloads to a single region in Google Cloud to simplify management and reduce latency. Implement basic security measures, such as firewalls and intrusion detection systems, to protect against common threats. Use a combination of managed services and self-managed services to balance cost and flexibility.

B. Design a multi-zone architecture within a single region to reduce costs while maintaining high availability. Implement security measures based on industry best practices, such as using strong passwords and multi-factor authentication. Use primarily self-managed services to have greater control over the environment.

C. Design a multi-region architecture with active-passive failover for high availability and disaster recovery. Implement appropriate security measures, such as data encryption at rest and in transit, to protect sensitive customer data. Use managed services whenever possible to reduce operational overhead and costs.

D. Defer considerations for high availability and disaster recovery until a later phase to minimize initial costs and complexity. Implement security measures as needed based on the sensitivity of the data. Use primarily open-source tools and technologies to minimize licensing costs.

Diagnostic Question Discussion

You work for a large financial institution that is planning a multi-phase migration of its on-premises workloads to Google Cloud. The migration involves sensitive customer data and requires high availability and disaster recovery capabilities. You want to design a cloud solution architecture that meets these requirements while minimizing costs and ensuring compliance with industry regulations.

What should you do?

A. Migrate all workloads to a single region in Google Cloud to simplify management and reduce latency. Implement basic security measures, such as firewalls and intrusion detection systems, to protect against common threats. Use a combination of managed services and self-managed services to balance cost and flexibility.

B. Design a multi-zone architecture within a single region to reduce costs while maintaining high availability. Implement security measures based on industry best practices, such as using strong passwords and multi-factor authentication. Use primarily self-managed services to have greater control over the environment.

C. Design a multi-region architecture with active-passive failover for high availability and disaster recovery. Implement appropriate security measures, such as data encryption at rest and in transit, to protect sensitive customer data. Use managed services whenever possible to reduce operational overhead and costs.

D. Defer considerations for high availability and disaster recovery until a later phase to minimize initial costs and complexity. Implement security measures as needed based on the sensitivity of the data. Use primarily open-source tools and technologies to minimize licensing costs.

Diagnostic Question Discussion

You are a Professional Cloud Architect working with a large retail customer that has a monolithic e-commerce application hosted on-premises. They are experiencing challenges with scalability and performance, especially during peak shopping seasons. They want to migrate this application to Google Cloud and modernize it to be more resilient, scalable, and cost-effective.

Your goal is to design a solution that meets their requirements.

What should you do? (choose two)

A. Migrate the application to Google Cloud using a phased approach, starting with a lift-and-shift migration and gradually modernizing components.

B. Deploy the entire application to a single, large Compute Engine instance to ensure resource availability and minimize management overhead.

C. Continue running the application on-premises and use Cloud CDN and Cloud Load Balancing to enhance performance and scalability.

D. Decompose the monolithic application into microservices and leverage managed services like Google Kubernetes Engine (GKE) and Cloud SQL.

E. Refactor the entire application to serverless architecture using Cloud Functions and Cloud Run to minimize operational overhead and maximize cost savings.

Diagnostic Question Discussion

You are a Professional Cloud Architect working with a large retail customer that has a monolithic e-commerce application hosted on-premises. They are experiencing challenges with scalability and performance, especially during peak shopping seasons. They want to migrate this application to Google Cloud and modernize it to be more resilient, scalable, and cost-effective.

Your goal is to design a solution that meets their requirements.

What should you do? (choose two)

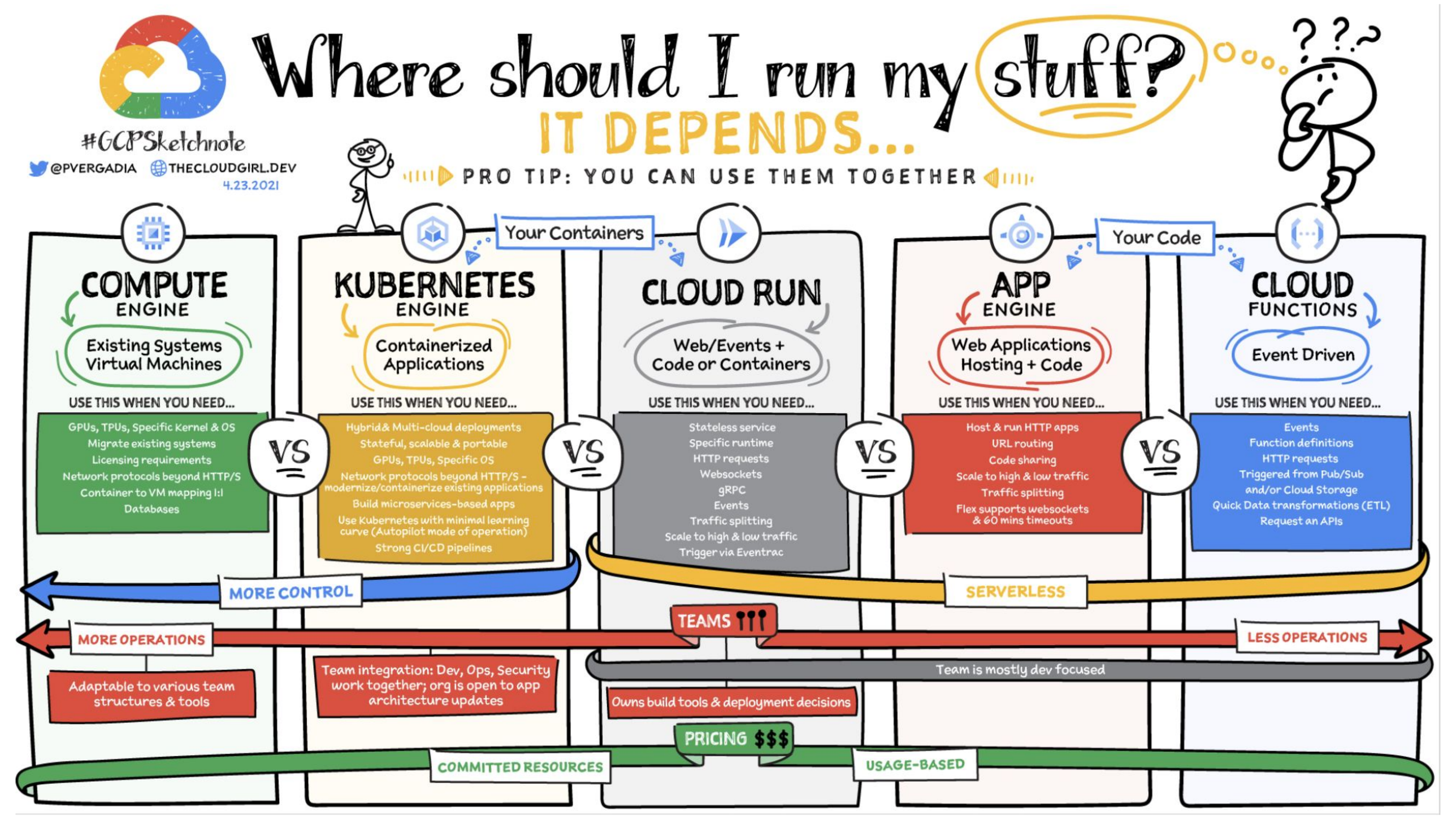
A. Migrate the application to Google Cloud using a phased approach, starting with a lift-and-shift migration and gradually modernizing components.

B. Deploy the entire application to a single, large Compute Engine instance to ensure resource availability and minimize management overhead.

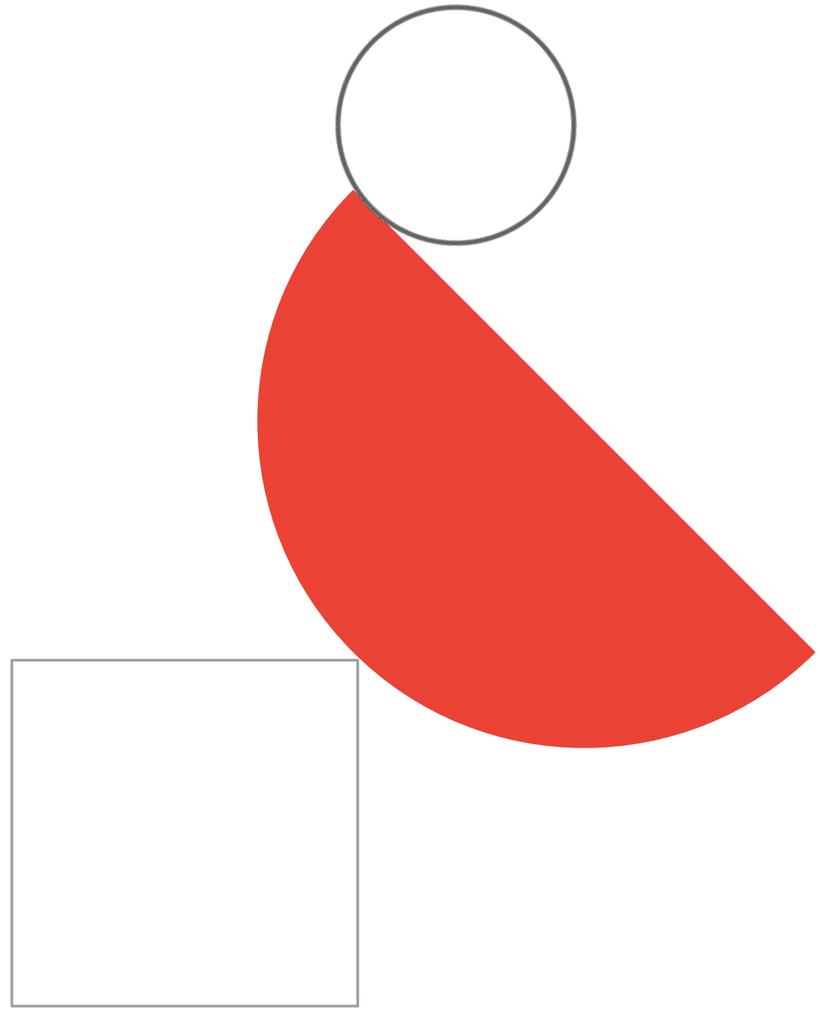
C. Continue running the application on-premises and use Cloud CDN and Cloud Load Balancing to enhance performance and scalability.

D. Decompose the monolithic application into microservices and leverage managed services like Google Kubernetes Engine (GKE) and Cloud SQL.

E. Refactor the entire application to serverless architecture using Cloud Functions and Cloud Run to minimize operational overhead and maximize cost savings.


#GCPSketchnote
@PVERGADIA
THECLOUDGIRL.DEV
4.23.2021
Where should I run my stuff?
IT DEPENDS...
PRO TIP: YOU CAN USE THEM TOGETHER
Your Containers
Your Code
COMPUTE ENGINE
Existing Systems Virtual Machines
USE THIS WHEN YOU NEED...
GPUs, TPUs, Specific Kernel & OS
Migrate existing systems
Licensing requirements
Network protocols beyond HTTP/S
Container to VM mapping 1:1
Databases
VS
KUBERNETES ENGINE
Containerized Applications
USE THIS WHEN YOU NEED...
Hybrid& Multi-cloud deployments
Stateful, scalable & portable
GPUs, TPUs, Specific OS
Network protocols beyond HTTP/S - modernize/containerize existing applications
Build microservices-based apps
Use Kubernetes with minimal learning curve (Autopilot mode of operation)
Strong CI/CD pipelines
VS
CLOUD RUN
Web/Events + Code or Containers
USE THIS WHEN YOU NEED...
Stateless service
Specific runtime
HTTP requests
Websockets
gRPC
Events
Traffic splitting
Scale to high & low traffic
Trigger via Eventrac
VS
APP ENGINE
Web Applications Hosting + Code
USE THIS WHEN YOU NEED...
Host & run HTTP apps
URL routing
Code sharing
Scale to high & low traffic
Traffic splitting
Flex supports websockets & 60 mins timeouts
VS
CLOUD FUNCTIONS
Event Driven
USE THIS WHEN YOU NEED...
Events
Function definitions
HTTP requests
Triggered from Pub/Sub and/or Cloud Storage
Quick Data transformations (ETL)
Request an APIs
MORE CONTROL
SERVERLESS
MORE OPERATIONS
LESS OPERATIONS
TEAMS
Adaptable to various team structures & tools
Team integration: Dev, Ops, Security work together; org is open to app architecture updates
Owns build tools & deployment decisions
Team is mostly dev focused
PRICING $$$
COMMITTED RESOURCES
USAGE-BASED

Google Compute Engine (GCE)



Google Compute Engine

Infrastructure as a Service (IaaS)

- vCPUs (cores) and Memory (RAM)
- Persistent disks
- Networking
- Linux or Windows

Exam Tips: *GCE is a basic IaaS service, but there are lots of details you're expected to know:*

- *Differences between PD images / snapshots / VM images.*
- *How to troubleshoot VM not booting up properly*
- *Custom image vs public image + startup scripts*
- *VM price differ between regions*
- *PDs are network-attach devices and - as such - consume VM bandwidth.*
- *VM network performance scales with # of vCPUs.*
- *etc…*

Compute Engine - how to differentiate between families?

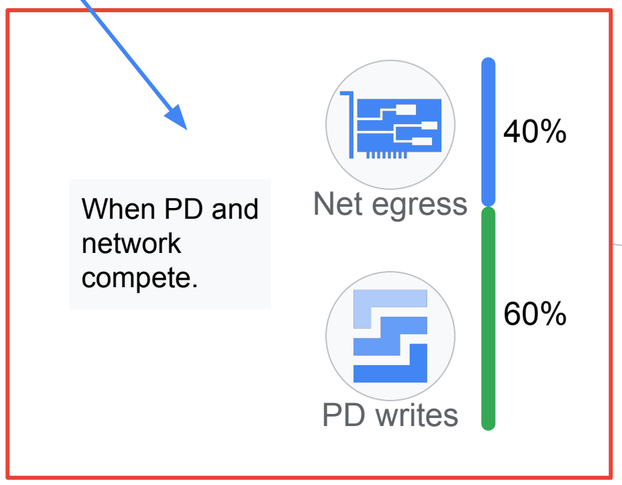
Best TCO	Balanced	Scale-out Optimized	Workload-Optimized		
•Web Serving •Steady-state LOB apps •Dev & Test environments •Small prod environments	•Enterprise apps •Medium databases •Web & App Serving	•Scale-out Workloads •Web Serving •Containerized microservices	•EDA •HPC •Scientific Modeling •AAA Gaming	•SAP HANA •Largest in memory DBs •Real-time data analytics •In-memory cache	•ML •HPC •Massive parallelized computation
Cost savings a priority	Leading perf and perf/$	Best Perf/$ for scale out workloads	Highest performance CPUs	Most memory on Compute Engine	Highest performance GPUs
Cost-Optimized (E2)	**General Purpose** (N2 and N2D)	**ScaleOut optimized Tau (T2D, T2A)**	**Compute-Optimized** (C2, C2D)	**Memory-Optimized** (M1, M2, M3)	**Accelerator-Optimized** (A2)

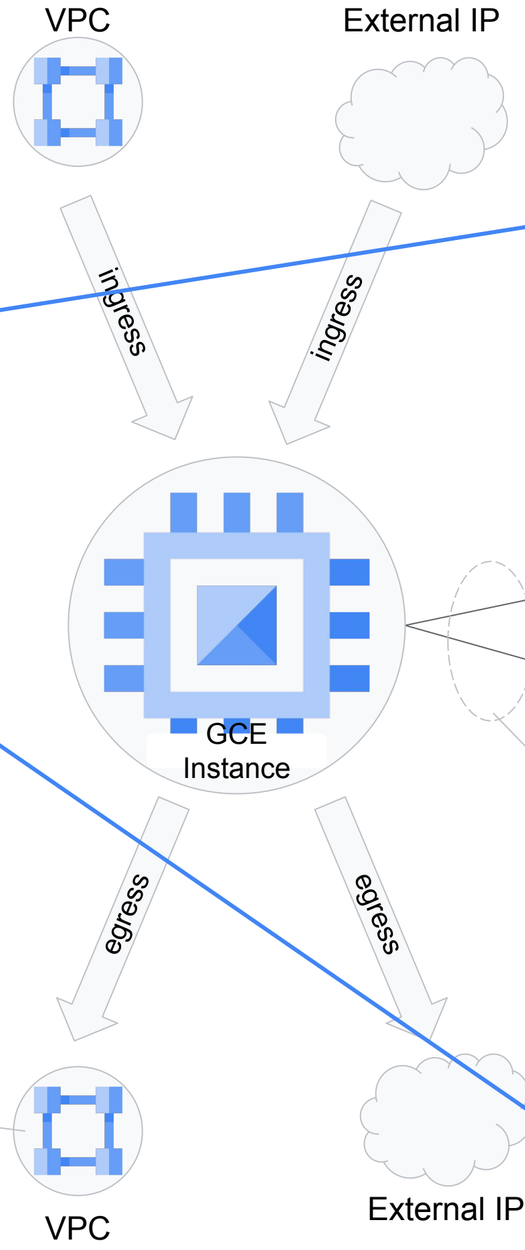
Compute Engine

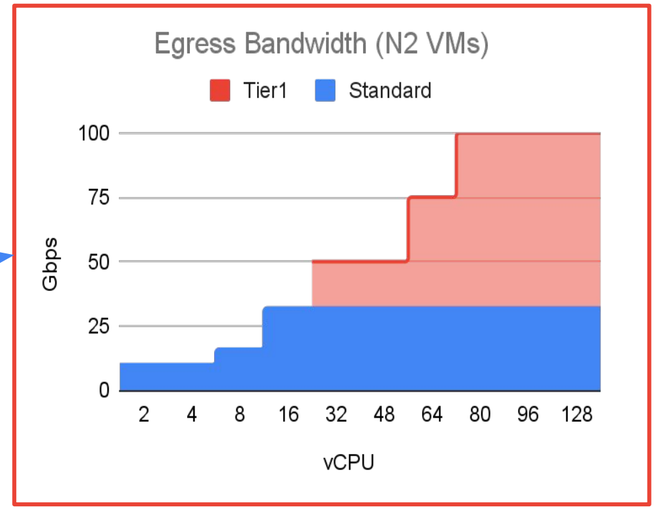
Network perspective

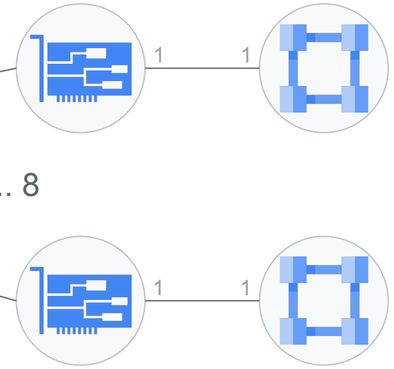
Exam Tips:

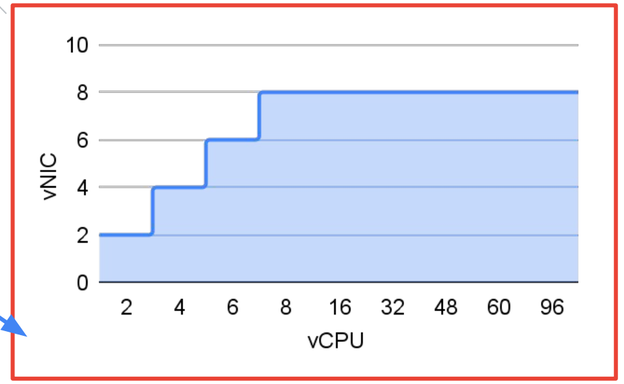
- *Network bandwidth limited & dependent on vCPU count (up to ~32Gbps for N2s + Tier1 extends further)*
- *You can expect the best network performance for traffic within the same zone, using internal IP addresses.*
- *Remember about multi-NIC VMs (up to 8)*
- *Storage is a network resource! => Network bandwidth shared between network AND disk activity*







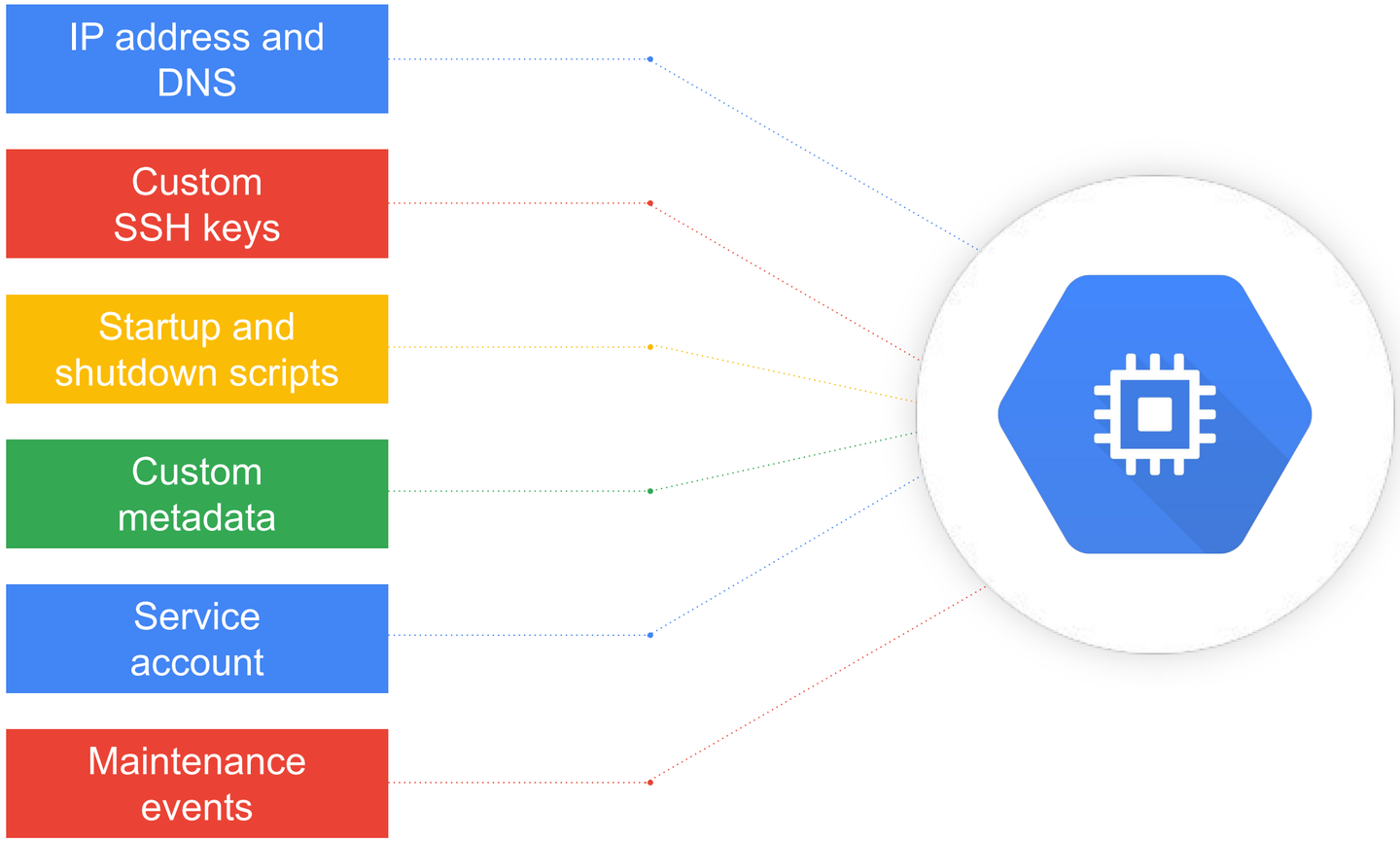




Compute Engine: Metadata Server

The metadata server stores information about the instance or project.

- Metadata request/response never leaves the physical host.
- Metadata information is encrypted on the way to the virtual machine host.
- Metadata server can generate a signed token for apps to verify the instance identity.



Compute Engine: Spot (Preemptible) VMs

Made for batch, fault-tolerant, and high throughput computing

Super-low-cost, short-term instances

- Up to 91% less than standard instances
- No maximum duration, may be preempted with 30-seconds notice (preemptible: max 24 hrs)
- Simple to use with graceful termination

Exam Tips:

- *Those use-cases usually pop up at the exam with regards to Spot VMs / Preemptibles.*
- *Can also be used in GKE clusters!*

Ideal for a variety of stateless, fault-tolerant workloads

- Genomics, pharmaceuticals
- Physics, math, computational chemistry
- Data processing (for example, with Hadoop or Cloud DataProc)
- Image handling, rendering, and media transcoding
- Monte Carlo simulations
- Financial services

Compute Engine: automate start & stop activities

Executed from metadata, either directly or from file:

- Startup:
  - gcloud compute instances create VM_NAME \
    --image-project=debian-cloud \
    --image-family=debian-10 \
    --metadata=startup-script='#! /bin/bash
    apt update
    apt -y install apache2
    cat <<EOF > /var/www/html/index.html
    <html><body> Linux startup script added directly. </body></html>
    EOF'

- Shutdown:
  - gcloud compute instances create example-instance
    --metadata-from-file=shutdown-script=FILE_PATH

To see output of startup/shutdown script:

- gcloud compute instances create example-instance --metadata
  shutdown-script="#! /bin/bash
  > # Shuts down Apache server
  > /etc/init.d/apache2 stop"

Exam Tips:

- *Startup / shutdown scripts are best-effort only!*
- *Startup / shutdown scripts are always run by root (Linux) / System (Windows)*
- *Shutdown scripts are especially useful for:*
  - *MIGs (to copy back processed data or logs before a VM goes down).*
  - *Spot / Preemptible VMs, which are much more vulnerable to be stopped.*
- *Startup / shutdown scripts can be set on VM or project (!!!) level -> will trigger for every VM. VM-level always take precedence (if exists, project-level script is not executed)*
- *Shutdown scripts have timeouts:*
  - *90s for standard instances*
  - *30s for Spot / Preemptible instances*

Diagnostic Question Discussion

You want to create a number of spot Linux virtual machine instances using Google Compute Engine. You want to properly shut down your application before the virtual machines are preempted.

What should you do?

A. Create a shutdown script named k99.shutdown in the /etc/rc.6.d/ director

B. Create a shutdown script registered as a xinetd service in Linux and configure a Cloud Monitoring endpoint check to call the servic

C. Create a shutdown script and use it as the value for a new metadata entry with the key shutdown-script in the Cloud Platform Console when you create the new virtual machine instance

D. Create a shutdown script, registered as a xinetd service in Linux, and use the gcloud compute instances add-metadata command to specify the service URL as the value for a new metadata entry with the key shutdown-script-url

Diagnostic Question Discussion

You want to create a number of spot Linux virtual machine instances using Google Compute Engine. You want to properly shut down your application before the virtual machines are preempted.

What should you do?

A. Create a shutdown script named k99.shutdown in the /etc/rc.6.d/ director

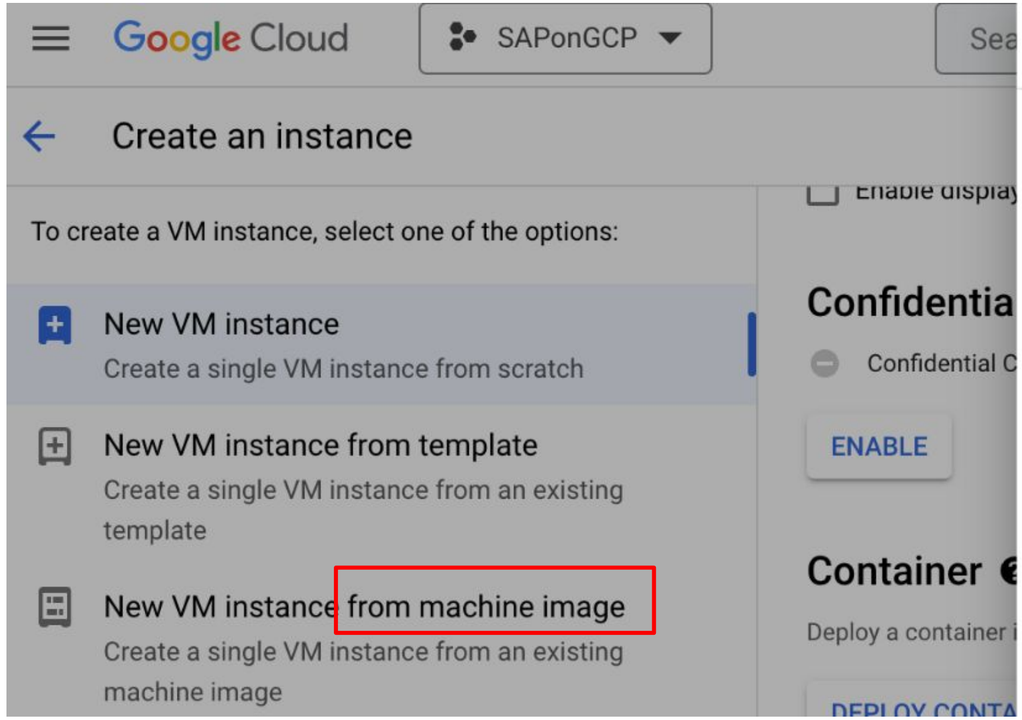
B. Create a shutdown script registered as a xinetd service in Linux and configure a Cloud Monitoring endpoint check to call the servic

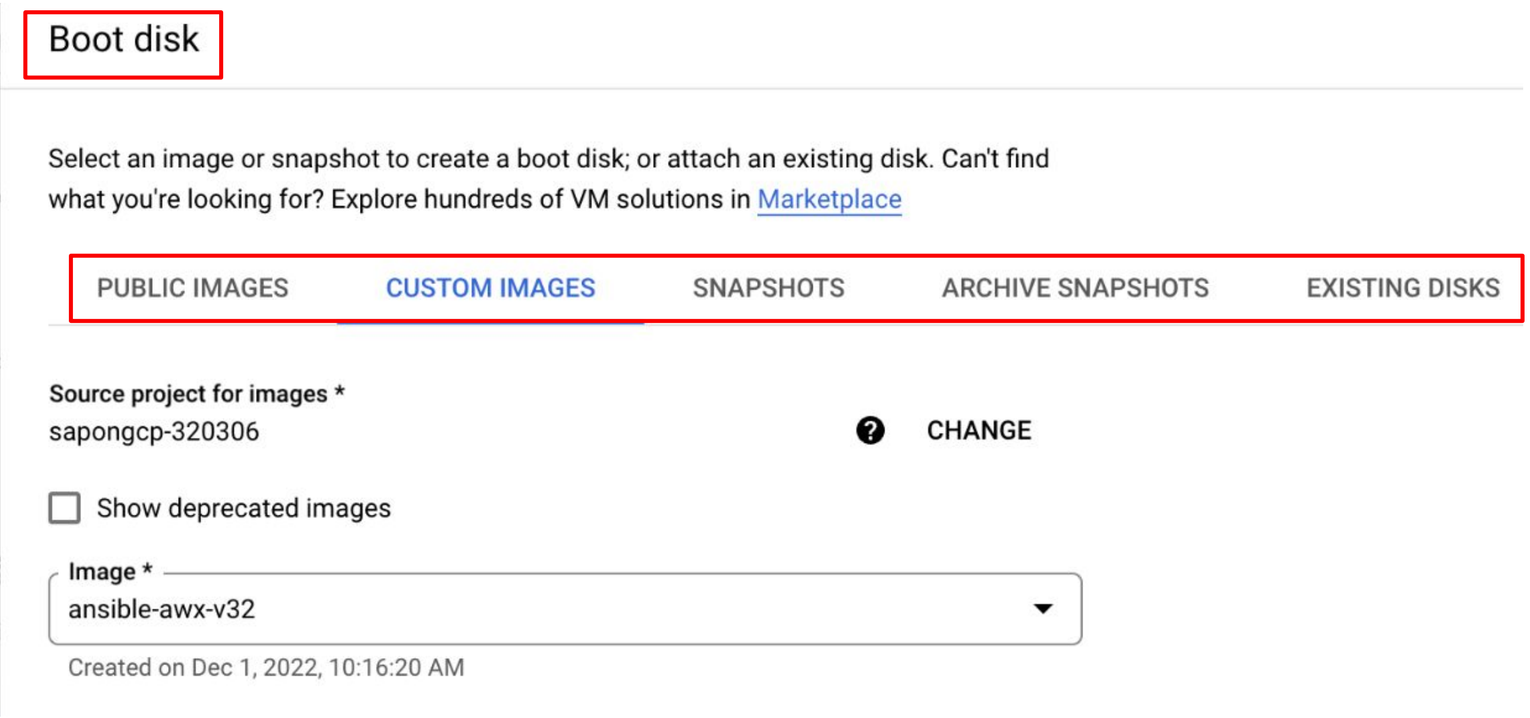
C. Create a shutdown script and use it as the value for a new metadata entry with the key shutdown-script in the Cloud Platform Console when you create the new virtual machine instance

D. Create a shutdown script, registered as a xinetd service in Linux, and use the gcloud compute instances add-metadata command to specify the service URL as the value for a new metadata entry with the key shutdown-script-url

Compute Engine creation

public OS image vs custom OS image vs snapshot vs machine image





Exam Tips:

- *Custom images should be centralized and controlled from lifecycle perspective (know what are image families and image states*
- *Public / Custom OS image IS NOT the same as "machine image"*
- *You can create a VM based on all of those options (public / custom OS image, snapshot, existing disk, machine image)*
- *You can 'automate' post-processing with startup script, regardless of how boot disk was created.*

Diagnostic Question Discussion

You need to deploy an application on Google Cloud that must run on a Debian Linux environment. The application requires extensive configuration in order to operate correctly. You want to ensure that you can install Debian distribution updates with minimal manual intervention whenever they become available.

What should you do?

A. Create a Compute Engine instance template using the most recent Debian image. Create an instance from this template, and install and configure the application as part of the startup script. Repeat this process whenever a new Google-managed Debian image becomes available.

B. Create a Debian-based Compute Engine instance, install and configure the application, and use OS patch management to install available updates.

C. Create an instance with the latest available Debian image. Connect to the instance via SSH, and install and configure the application on the instance. Repeat this process whenever a new Google-managed Debian image becomes available.

D. Create a Docker container with Debian as the base image. Install and configure the application as part of the Docker image creation process. Host the container on Google Kubernetes Engine and restart the container whenever a new update is available.

Diagnostic Question Discussion

You need to deploy an application on Google Cloud that must run on a Debian Linux environment. The application requires extensive configuration in order to operate correctly. You want to ensure that you can install Debian distribution updates with minimal manual intervention whenever they become available.

What should you do?

A. Create a Compute Engine instance template using the most recent Debian image. Create an instance from this template, and install and configure the application as part of the startup script. Repeat this process whenever a new Google-managed Debian image becomes available.

B. Create a Debian-based Compute Engine instance, install and configure the application, and use OS patch management to install available updates.

C. Create an instance with the latest available Debian image. Connect to the instance via SSH, and install and configure the application on the instance. Repeat this process whenever a new Google-managed Debian image becomes available.

D. Create a Docker container with Debian as the base image. Install and configure the application as part of the Docker image creation process. Host the container on Google Kubernetes Engine and restart the container whenever a new update is available.

More about patch management here.

Shielded VMs

Exam Tips: *Using Shielded VMs is a best practice in GCP!*

Secure Boot	vTPM	Integrity Monitoring	Result/implications
ON	ON	ON	Most secure. Allows for use of vTPM for data encryption using vTPM protected key, Secure Boot to prevent malicious rootkits and bootkits, and Integrity Monitoring to alert to any changes in boot process. Secure Boot may not be compatible with customers drivers or other software.
OFF	ON	ON	Default when creating a GCP VM. Allows for use of vTPM for data encryption using vTPM protected key and Integrity Monitoring to alert to any changes in boot process. If customer has unsigned drivers or low level software this is the most secure option as Secure Boot would not be compatible.
OFF	OFF	OFF	Least secure. No benefits of Shielded VM. This is not recommended.

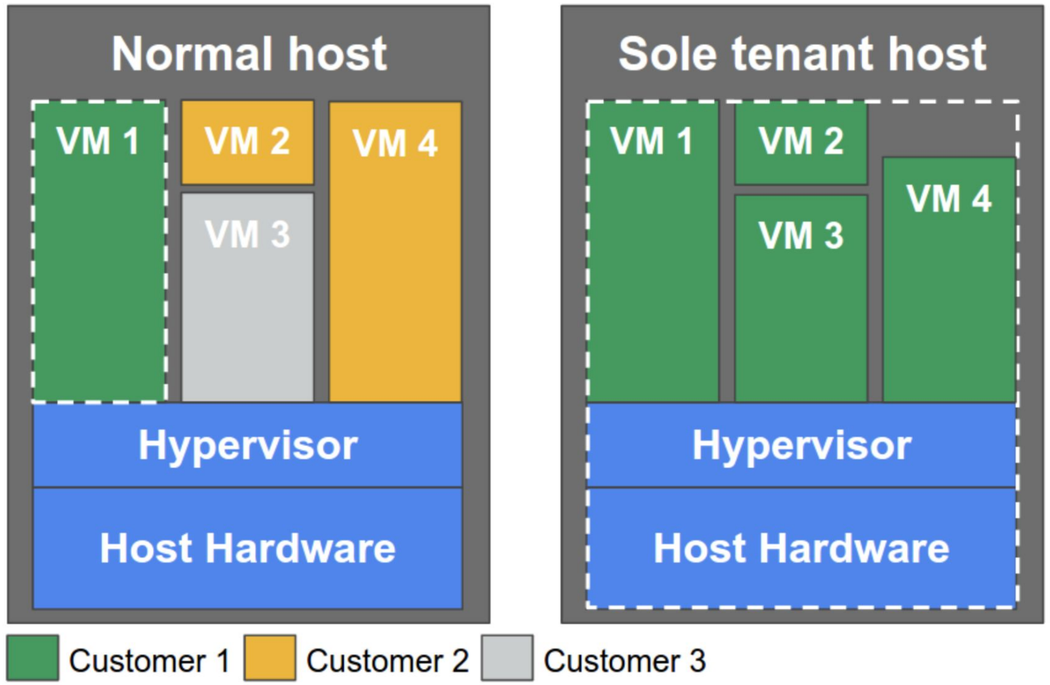
```
gcloud compute instances update instance name \
```

Feature	Flag to Turn On	Flag to Turn Off
Secure boot	--shielded-secure-boot	--no-shielded-secure-boot
vTPM (measure boot)	--shielded-vtpm	--no-shielded-vtpm
Integrity monitoring	--shielded-integrity-monitoring	--no-shielded-integrity-monitoring

Sole-Tenant Nodes

Regular VMs on regular machines, dedicated specifically to your workloads.

- Dedicated hardware
- Mix-and-match VMs to consume host resources
- Full access to host resources for 10% premium*

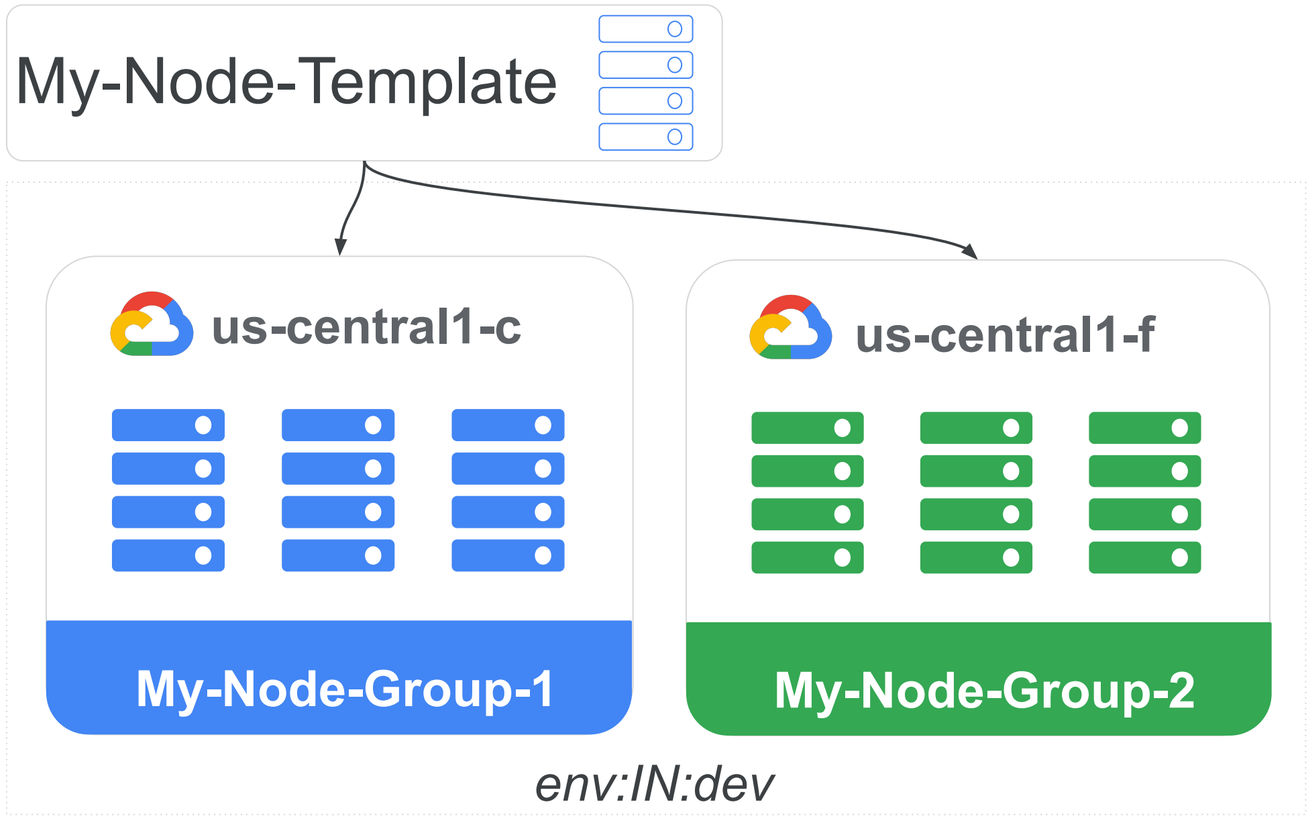


*10% Premium based on on-demand price

Quick Start for Sole-Tenant Nodes (1/2)

Each Sole-Tenant Node has a 1:1 mapping to a physical Host and represents a reserved host.

Step 1: Reserve Sole-Tenant Node(s)




```
// 1. CREATE NODE TEMPLATE
$ gcloud compute sole-tenancy \
node-templates create my-node-template \
--node-type n1-node-96-624 \
--region us-central1
```

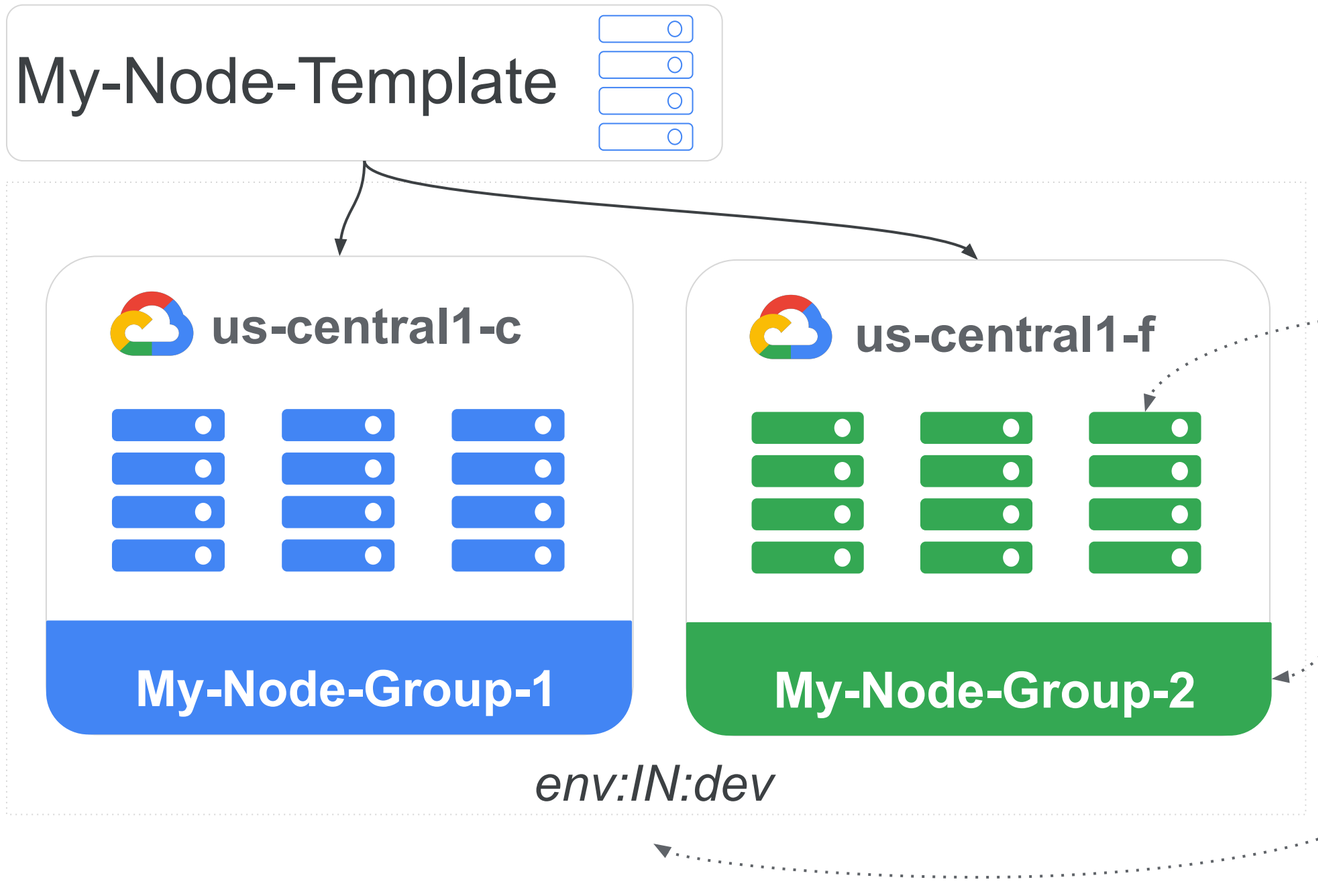
```
// 2. CREATE NODE GROUP OF 3 NODES
$ gcloud compute sole-tenancy \
node-groups create my-node-group-1 \
--node-template my-node-template \
--target-size 3 \
--zone us-central1-c
```

```
// 2b. [FOR ILLUSTRATION] CREATE ANOTHER
$ gcloud compute sole-tenancy \
node-groups create my-node-group-2 \
--node-template my-node-template \
--target-size 3 \
--zone us-central1-f
```

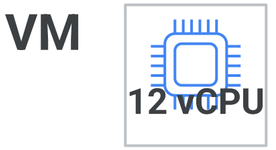
Quick Start for Sole-Tenant Nodes (2/2)

Exam Tips: *More info on provisioning VMs on sole-tenant nodes can be found here.*

Step 1: Reserve Sole-Tenant Node(s)



2. Schedule Instance(s)



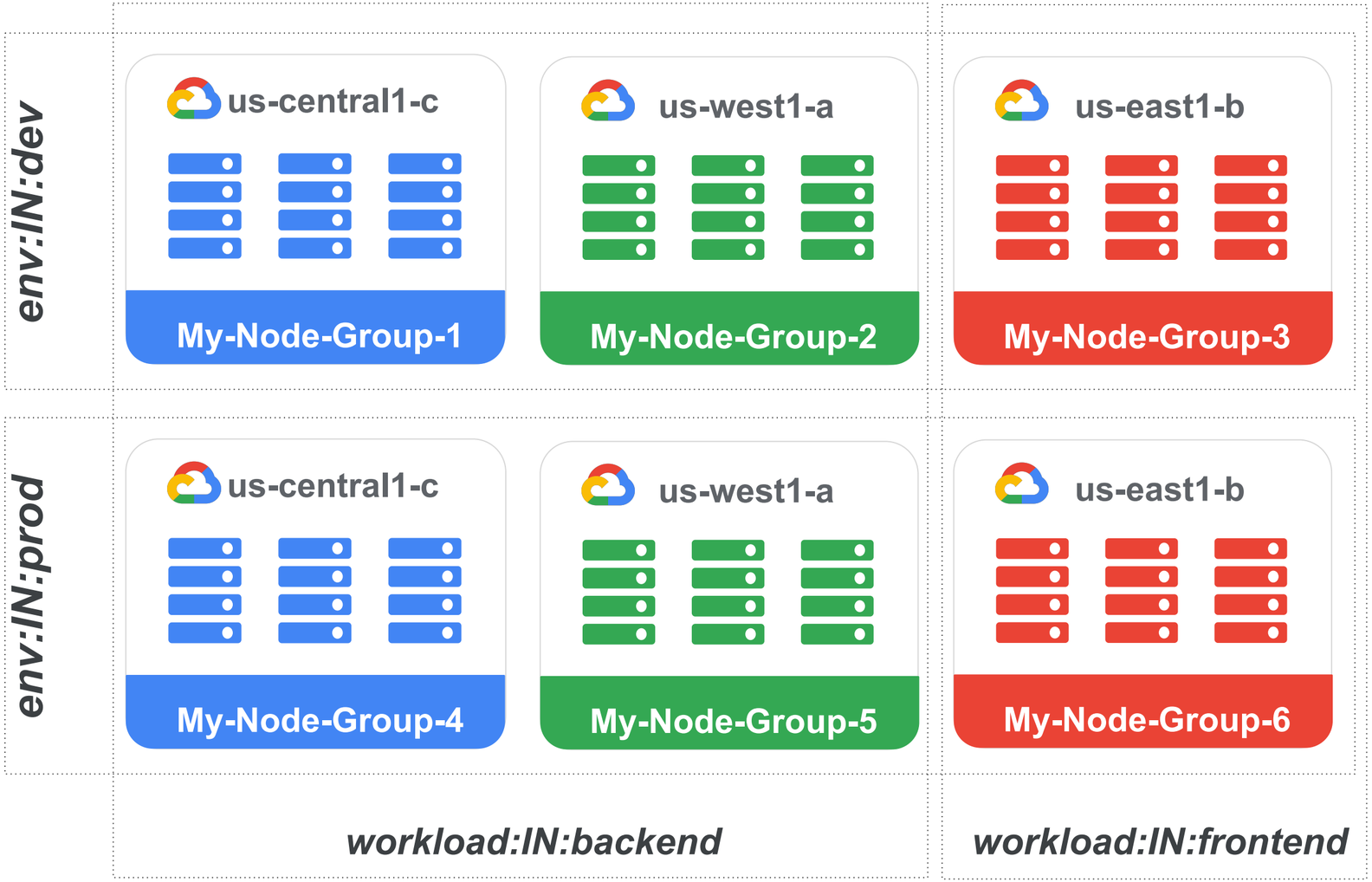
3 ways to schedule:

```
// SCHEDULE ONTO A SPECIFIC NODE
$ gcloud compute instances create \
INSTANCE_NAME --node=NODE_NAME
```

```
// SCHEDULE ON ANY NODE IN NODE GROUP
$ gcloud compute instances create \
INSTANCE_NAME \
--node-group=NODE_GROUP_NAME
```

```
// SCHEDULE ONTO ANY HOST WITH
// MATCHING LABELS
$ gcloud compute instances create
INSTANCE_NAME --zone ZONE
--node-affinity-file=FILE.json
```

Sole-Tenant Nodes: Using Node Affinity Labels

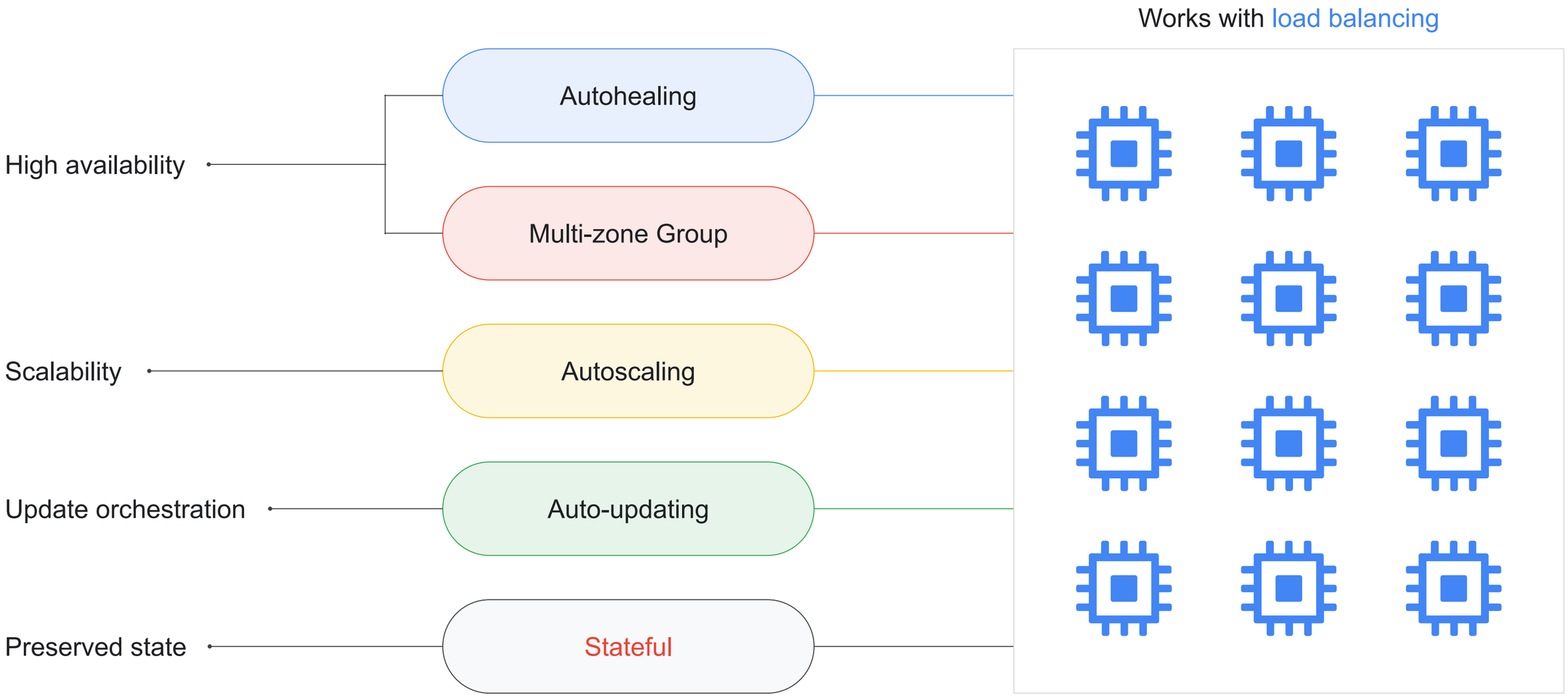



```
// SCHEDULE ONTO ANY HOST WITH
// MATCHING LABELS
$ gcloud compute instances create
INSTANCE_NAME --zone ZONE
--node-affinity-file=FILE.json
```

```
// SCHEDULE ONTO ANY HOST WITH
// IN MY-NODE-GROUP-1
$ gcloud compute instances create
INSTANCE_NAME --zone ZONE
--node-group my-node-group-1
```

Managed Instance Groups: Run VMs at Scale

Up to thousands of VMs



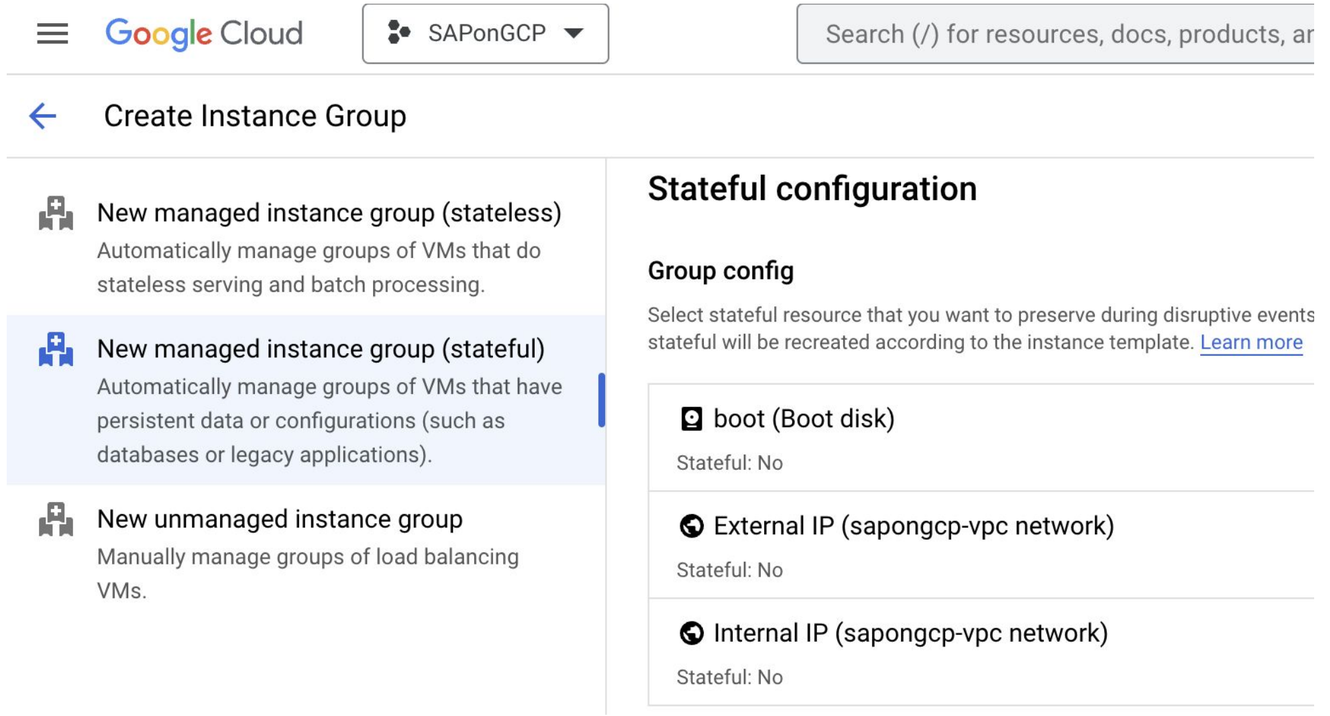
Exam Tips: *pros & cons of "ready" custom OS image vs public image + startup scripts*

Stateful vs stateless

And why stateless is usually preferred

Exam Tips:

- *Here a look at this document.*
- *Prefer stateless. Use stateful only when necessary, eg:*
  - *Databases*
  - *Data processing apps (Kafka etc)*
  - *Legacy monoliths*



Stateful	Stateless (PREFERRED!)
Each server retains information about it's client sessions, such as the current state of an application or the content of a user shopping cart.	Server does not retain any information about the client sessions. Each request made by the client is treated as an independent transaction, and the server does not maintain any memory of previous requests.
Not perfect if we have multiple backends that can serve the requests…	Greater scalability and flexibility.
Can scale up easily. Can't scale down easily since each server keeps its' state.	Ability to scale up & down easily

Choosing instance groups for Compute Engine

Type of Instance Group		Properties of Instances	Feature
Unmanaged		Heterogeneous	
Managed		Homogeneous	Instance Templates Autoscaling
	Zonal	Same zone	Latency consistency
	Regional	Different zones	Reliability

Exam Tip:

- *Unmanaged are used to group EXISTING, different VMs under one "umbrella" and balance traffic to healthy ones only. For example, used in lift&shift migrations.*
- *You can't update existing instance template (need to create a new one)*
- *Know the difference between scale-out and scale-up!*

```
gcloud compute instance-groups managed create [*INSTANCE_GROUP_NAME*] \
    --size=[*SIZE*] \
    --template=[*INSTANCE_TEMPLATE_NAME*] \
    --zone=[*ZONE*]
```

MIG - Autoscaling

CPU Utilization

Treats the target CPU utilization level as a fraction of the average use of all vCPUs over time in the instance group

External HTTPS Capacity

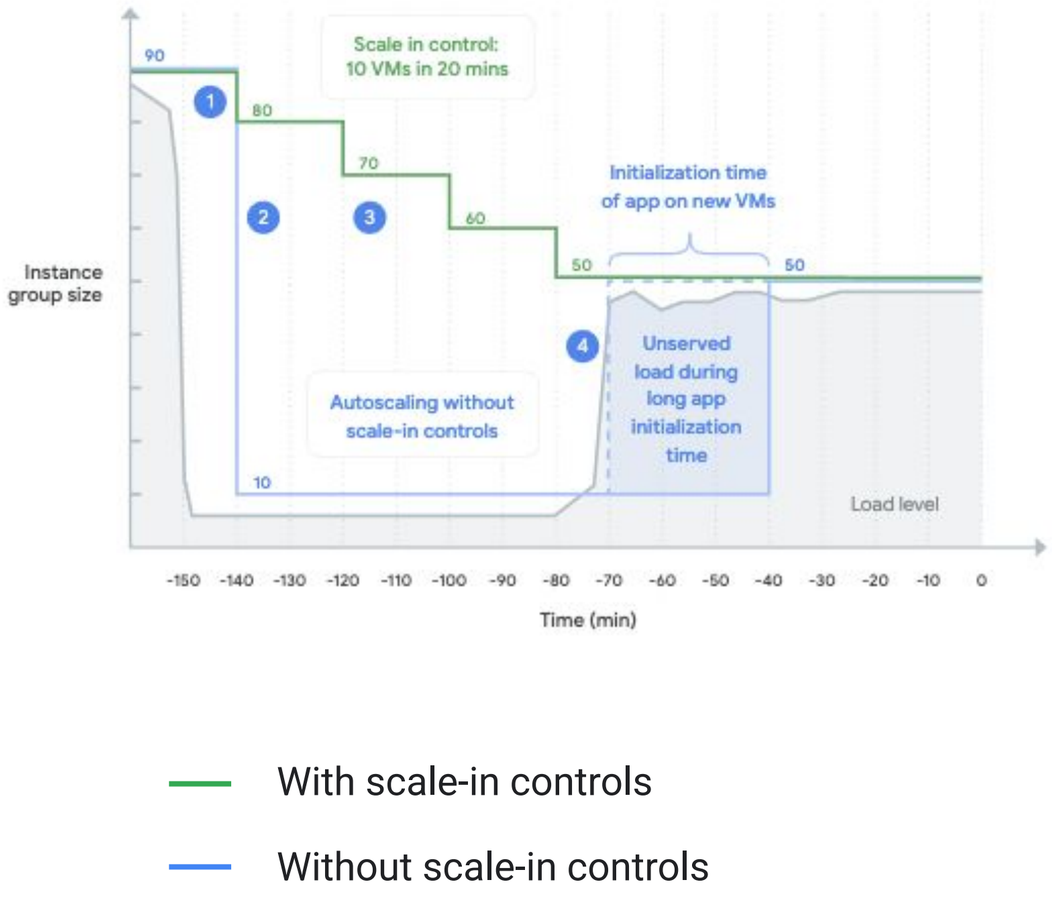
Autoscaling works with maximum backend utilization and maximum requests per second/instance

Cloud Monitoring Metrics

Per Instance or Per Group
Standard or custom metrics
Not for log-based metrics

Schedules

Additional autoscaler
Up to 128 schedules
Min instances
Duration
Start time & Recurrence



Diagnostic Question Discussion

Your company is running its application workloads on Compute Engine. The applications have been deployed in production, acceptance, and development environments. The production environment is business-critical and is used 24/7, while the acceptance and development environments are only critical during office hours. Your CFO has asked you to optimize these environments to achieve cost savings during idle times.

What should you do?

A. Create a shell script that uses the gcloud command to change the machine type of the development and acceptance instances to a smaller machine type outside of office hours. Schedule the shell script on one of the production instances to automate the task.

B. Use Cloud Scheduler to trigger a Cloud Function that will stop the development and acceptance environments after office hours and start them just before office hours.

C. Deploy the applications using managed instance groups and enable autoscaling based on appropriate metrics.

D. Use regular Compute Engine instances for the production environment, and use spot VMs for the acceptance and development environments.

Diagnostic Question Discussion

Your company is running its application workloads on Compute Engine. The applications have been deployed in production, acceptance, and development environments. The production environment is business-critical and is used 24/7, while the acceptance and development environments are only critical during office hours. Your CFO has asked you to optimize these environments to achieve cost savings during idle times.

What should you do?

A. Create a shell script that uses the gcloud command to change the machine type of the development and acceptance instances to a smaller machine type outside of office hours. Schedule the shell script on one of the production instances to automate the task.

B. Use Cloud Scheduler to trigger a Cloud Function that will stop the development and acceptance environments after office hours and start them just before office hours.

C. Deploy the applications using managed instance groups and enable autoscaling based on appropriate metrics.

D. Use regular Compute Engine instances for the production environment, and use spot VMs for the acceptance and development environments.

VM Pricing and cost optimization

Sustained Use Discounts (SUD)

Up to 30% savings on Compute Engine and Cloud SQL

Committed Use Discounts (CUD)

Up to 70% savings without upfront fees or instance-type lock-in

Spot / Preemptible VM instances

Up to 91% savings on workloads that can be interrupted, like data mining and data processing

Per second billing

Up to 38% savings by paying per second, not per hour

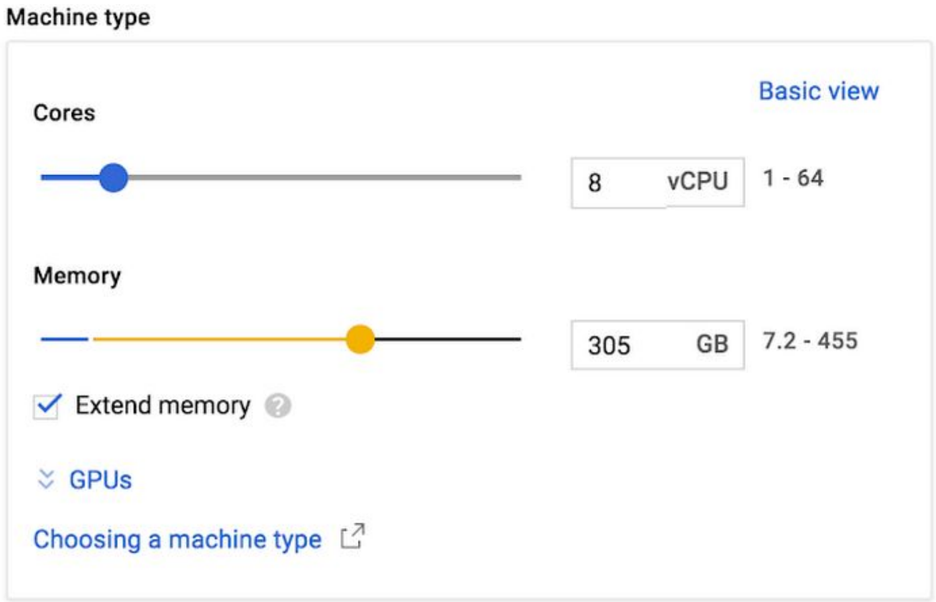
Network Service Tiers

Pick performance and get 70% more bandwidth than other clouds, or pick cost savings and save up to 9% compared to other clouds

Rightsizing (incl. choosing optimal GCE families) and Custom Machine Types

Exam Tips:

- Common pattern for optimization costs for unused PDs: you can create a snapshot, and delete the disk to reduce the maintenance cost of that disk by 35% to 92%.
- For premium OS, you're billed for license per vCPU per second.
- Bring Your Own License is an option for some OSes
- Use Extended memory to save on OS license costs.



Diagnostic Question Discussion

To reduce costs, the Director of Engineering has required all developers to move their development infrastructure resources from on-premises virtual machines (VMs) to Google Cloud Platform. These resources go through multiple start/stop events during the day and require state to persist. You have been asked to design the process of running a development environment in Google Cloud while providing cost visibility to the finance department.

Which two steps should you take? (Choose two.)

A. Use persistent disks to store the state. Start and stop the VM as needed

B. Use the --auto-delete flag on all persistent disks and terminate the VM

C. Apply VM CPU utilization label and include it in the BigQuery billing export

D. Use Google BigQuery billing export and labels to associate cost to groups

E. Store all state into local SSD, snapshot the persistent disks, and terminate the VM

F. Store all state in Google Cloud Storage, snapshot the persistent disks, and terminate the VM

Diagnostic Question Discussion

To reduce costs, the Director of Engineering has required all developers to move their development infrastructure resources from on-premises virtual machines (VMs) to Google Cloud Platform. These resources go through multiple start/stop events during the day and require state to persist. You have been asked to design the process of running a development environment in Google Cloud while providing cost visibility to the finance department.

Which two steps should you take? (Choose two.)

A. Use persistent disks to store the state. Start and stop the VM as needed

B. Use the --auto-delete flag on all persistent disks and terminate the VM

C. Apply VM CPU utilization label and include it in the BigQuery billing export

D. Use Google BigQuery billing export and labels to associate cost to groups

E. Store all state into local SSD, snapshot the persistent disks, and terminate the VM

F. Store all state in Google Cloud Storage, snapshot the persistent disks, and terminate the VM

Migrate for Compute Engine

Lift&Shift your VMWare, AWS, Azure workloads to GCE

- Purpose-built, enterprise-grade
- Migrate from on-prem or other clouds
- Proven at scale, having migrated customers w/ thousands of workloads
- Success across healthcare, energy, government, manufacturing, and more

Agentless

Nothing to install on source machines

Minimize complexity, reduce IT labor requirements by 5+ hours per server, keep migrations on track.

Streaming

Migrate storage while apps run in GCP

Eliminate long upfront data transfers and unpredictable maintenance windows, enabling fast time-to-cloud and reduced downtime.

Frictionless

Automate migration and in-cloud conversion

Reduce touch points for IT, provide uninterrupted experience for line of business owners and end users.

Diagnostic Question Discussion

You need to host an application on a Compute Engine instance in a project shared with other teams. You want to prevent the other teams from accidentally causing downtime on that application.

What should you do?

A. Use a Shielded VM.

B. Use a Spot VM.

C. Use a sole-tenant node.

D. Enable deletion protection on the instance.

Diagnostic Question Discussion

You need to host an application on a Compute Engine instance in a project shared with other teams. You want to prevent the other teams from accidentally causing downtime on that application.

What should you do?

A. Use a Shielded VM.

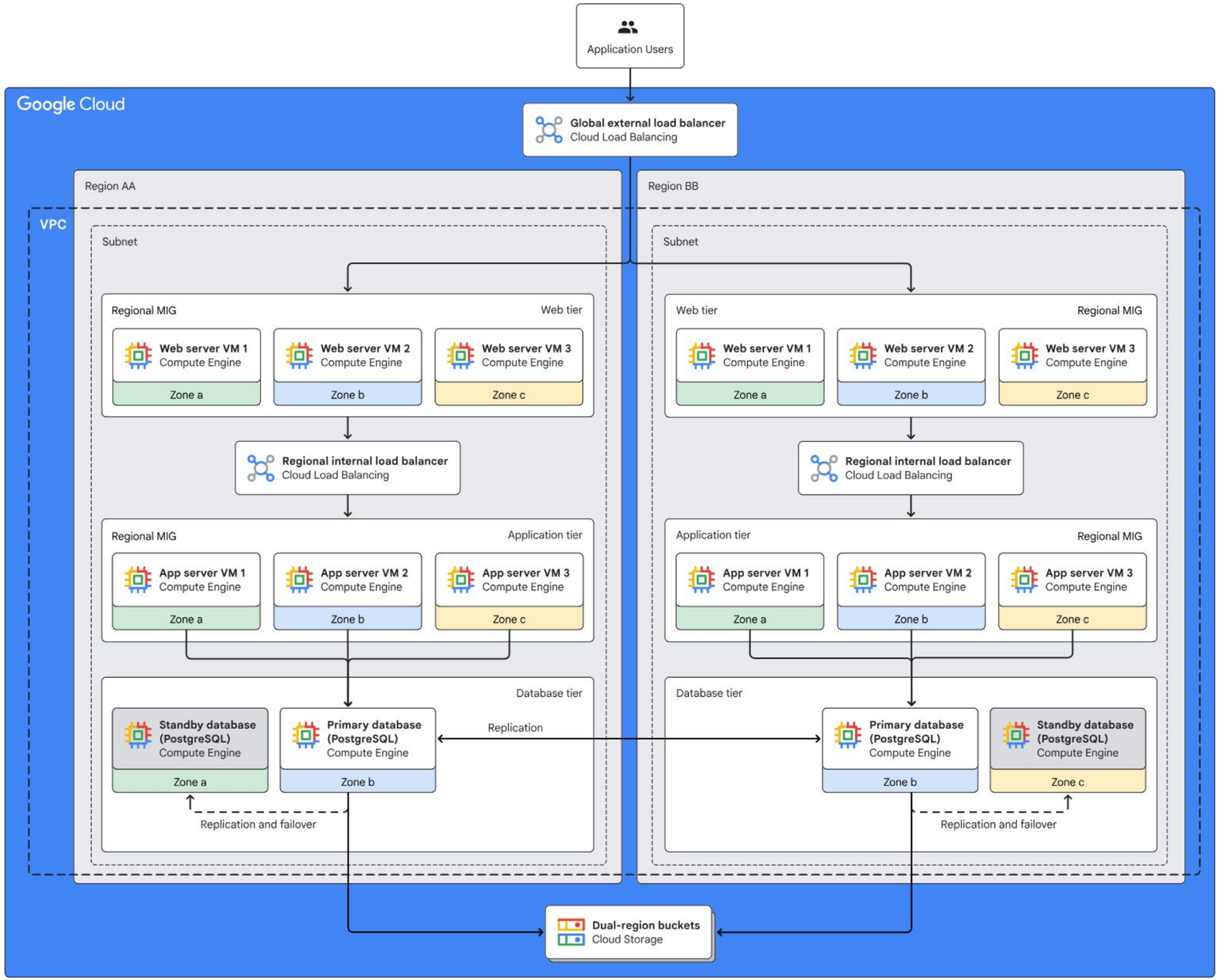
B. Use a Spot VM.

C. Use a sole-tenant node.

D. Enable deletion protection on the instance.

https://cloud.google.com/compute/docs/instances/preventing-accidental-vm-deletion

[Example] Multi-regional deployment on Compute Engine



Compute Engine

#GCPSketchnotes

@PVERGADIA THECLOUDGIRL.DEV

What is Compute Engine?

CUSTOMIZABLE VIRTUAL MACHINES IN GOOGLE CLOUD

THERE ARE 2 MACHINE TYPES

PREDEFINED
n1-standard-1 =
(1 vCPU, 3.75 GB memory)

CUSTOM
vCPU + Memory

CORES — 1 vCPU — 1-96

MEMORY — 3.75 GB — 1-6.5

THERE ARE 3 MACHINE TYPE FAMILIES

GENERAL PURPOSE Machine Type
- General Servers
- Websites
- Databases

COMPUTE OPTIMIZED Machine Type
- High Performance Computing
- Gaming
- Electronic Design Automation
- Single Threaded Applications

MEMORY OPTIMIZED Machine Type
- Large In-Memory Databases (SAP HANNA)
- In-Memory Analytics

Compute Engine Use case (example)

- FIG. 1 — Websites
- FIG. 2 — Databases
- FIG. 3 — Legacy Monolithic Apps
- FIG. 4 — Containers
- FIG. 5 — Windows Apps

Compute Engine PRICING

SUSTAINED USE SAVINGS
Automatic discounts for running VMs a significant portion of the month

COMMITTED USE DISCOUNT
Up to 57% savings with no up-front cost

PREEMPTIBLE VMs
Up to 80% savings and run batch jobs & fault-tolerant workloads

RIGHT SIZE RECOMMENDATIONS
Suggests resizing for efficiency and cost

How does it WORK??

CREATE

region
+ zone
+ machine type (cpu & memory)
= Instance

BACKUPS

- Automatic Scheduled snapshots
- Live migration — Keep apps running during maintenance

HIGH AVAILABILITY

- Automatic failover to another zone or region (ZONE A, AUTOMATIC FAILOVERS, ZONE B)
- MANAGED INSTANCE GROUPS (MIGs)
  - Autohealing
  - Load balancing
  - Automatic updates
  - Support for stateful workloads
  - Container support
- AUTOSCALER - 3 types of policies:
  1. CPU utilization = more than 60% → create new instance
  2. HTTP(S) load balancers service capacity — Requests per second or utilization
  3. Cloud monitoring metrics

Cloud Dataproc

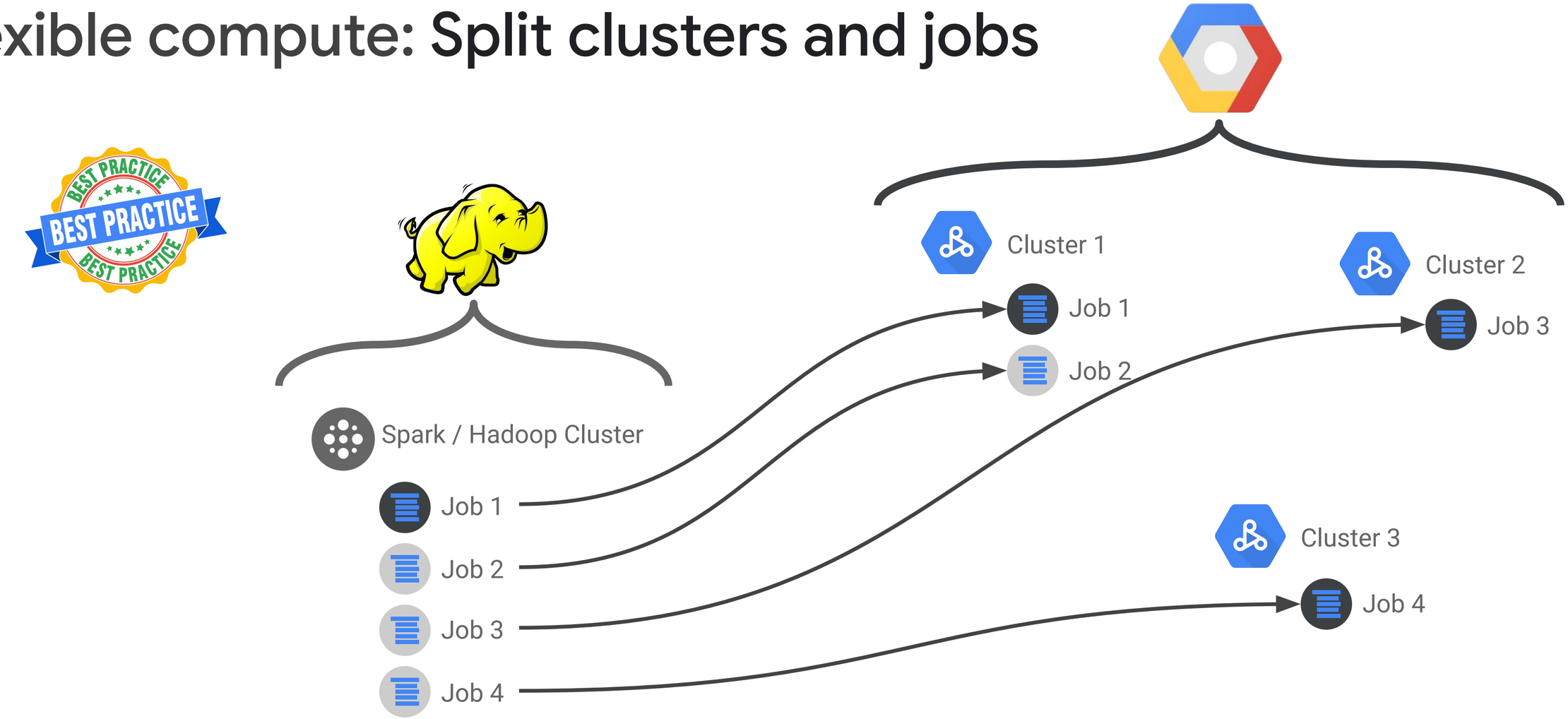
The benefits of Hadoop/Spark on Cloud

On premises	On compute engine	Cloud Dataproc
Custom code	Custom code	Custom code
Monitoring/Health	Monitoring/Health	Monitoring/Health
Dev integration	Dev integration	Dev integration
Scaling	Scaling	Scaling
Job submission	Job submission	Job submission
GCP connectivity	GCP connectivity	GCP connectivity
Deployment	Deployment	Deployment
Creation	Creation	Creation

Self-managed Google managed

Exam Tip: *if exam question mentions Apache Hadoop / Spark / Pig / Hive, plus it's clear that the customer already invested in building the pipelines in on-premises and does not want to lose it, you should probably go with Dataproc.*

Flexible compute: Split clusters and jobs



Exam Tips:

- *When thinking about **Dataproc**, you should really think about per-job, ephemeral, auto-scaling clusters with auto-shutdown after the task is completed.*
- *Using Spot/Preemptible VMs for **secondary** Dataproc workers is a common pattern.*
- *Switching from HDFS to GCS is also a best practice in most cases.*

Diagnostic Question Discussion

You need to migrate Hadoop jobs for your company's Data Science team without modifying the underlying infrastructure. You want to minimize costs and infrastructure management effort.

What should you do?

A. Create a Dataproc cluster using standard and spot worker instances.

B. Create a Dataproc cluster using spot worker instances only.

C. Manually deploy a Hadoop cluster on Compute Engine using standard instances.

D. Manually deploy a Hadoop cluster on Compute Engine using spot instances.

https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms

Diagnostic Question Discussion

You need to migrate Hadoop jobs for your company's Data Science team without modifying the underlying infrastructure. You want to minimize costs and infrastructure management effort.

What should you do?

A. **Create a Dataproc cluster using standard and spot worker instances.**

B. Create a Dataproc cluster using spot worker instances only.

C. Manually deploy a Hadoop cluster on Compute Engine using standard instances.

D. Manually deploy a Hadoop cluster on Compute Engine using spot instances.

https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms

Dataproc #GCPSketchnote

@PVERGADIA THECLOUDGIRL.DEV

12.07.2020

Managing Apache Hadoop cluster is hard!

ERIN

We need a tool that make it easy to gain insights from our data.

Yes, it is slow, difficult & expensive!

SAM

Dataproc!

How is a typical Hadoop/Spark cluster installation different from Dataproc?

Setup
Configure
Optimize
Utilize
Monitor
Shutdown

Typical Cluster Activities

90 seconds from start to finish

Dataproc

Simple deployment
Rapid development
Cost-effective use

Cluster Dataproc Activities

Familiar open source tools

Customizable machines
Easily autoscale up or down

Ephemeral clusters on-demand
Save cost

Tightly Integrated with
Google Cloud Platform

Rapid cluster creation
Spin up a cluster in 90 seconds

How do I move Spark/Hadoop jobs to Dataproc?

Copy Data → Update File Path → Go

GCS

from hdfs:// to gs://

Run your Spark, Hadoop, Hive, or Pig script

Can I right-size clusters per use case?

Storage
Cloud Storage

Hive Metastore
Cloud SQL

Dataproc	Dataproc	Dataproc	Dataproc
375GB local SSD	100GB PD	500GB PD-SSD	150GB PD
MapReduce ETL	Machine learning	Hive analytics	Business reporting

How do I use ephemeral & long-lived clusters to save cost?

Clusters per job

Cluster Dataproc
Cluster Dataproc
Cluster Dataproc
Cluster Dataproc

Client
Client
Client

Edge Nodes
Compute Engine

Storage
Cloud Storage

Semi-long-lived clusters - group and select by label

Production (1.2)
Prod 1 Dataproc
Prod 2 Dataproc

Development (Preview)
Dev cluster
Dataproc

Clients

How does it work?

Disaggregates storage & compute

External Applications
Application Logs

Data Sources
Storage Cloud Storage
Storage BigQuery
Storage Cloud Bigtable

Production Dataproc

External Applications
Test Dataproc
Development Dataproc

Cluster monitoring
Monitor Stackdriver
Logs Logging

Data Sources
Storage Cloud Storage
Storage BigQuery
Storage Cloud Bigtable

Data science
Analysis Notebook

Cloud Dataflow

Cloud Dataflow

The fully-managed, serverless, auto-optimized data processor that simplifies development and management of stream and batch pipelines

Exam Tips:

- *if exam question mentions Apache Beam -> most probably answer is Dataflow.*
- *When you're starting from scratch with ETL, Dataflow is preferred over Dataproc!*
- *KEY thing about Dataflow: it's able to serve BOTH batch and streaming within a SINGLE pipeline.*

Stream Analytics

- Works with Cloud Pub/Sub to deliver stream analytics
- Real-time data processing with "exactly-once" semantic

Unified with Batch

- No more Lambda architecture
- Apache Beam provides unified batch & streaming
- Reuse skills, tools and code

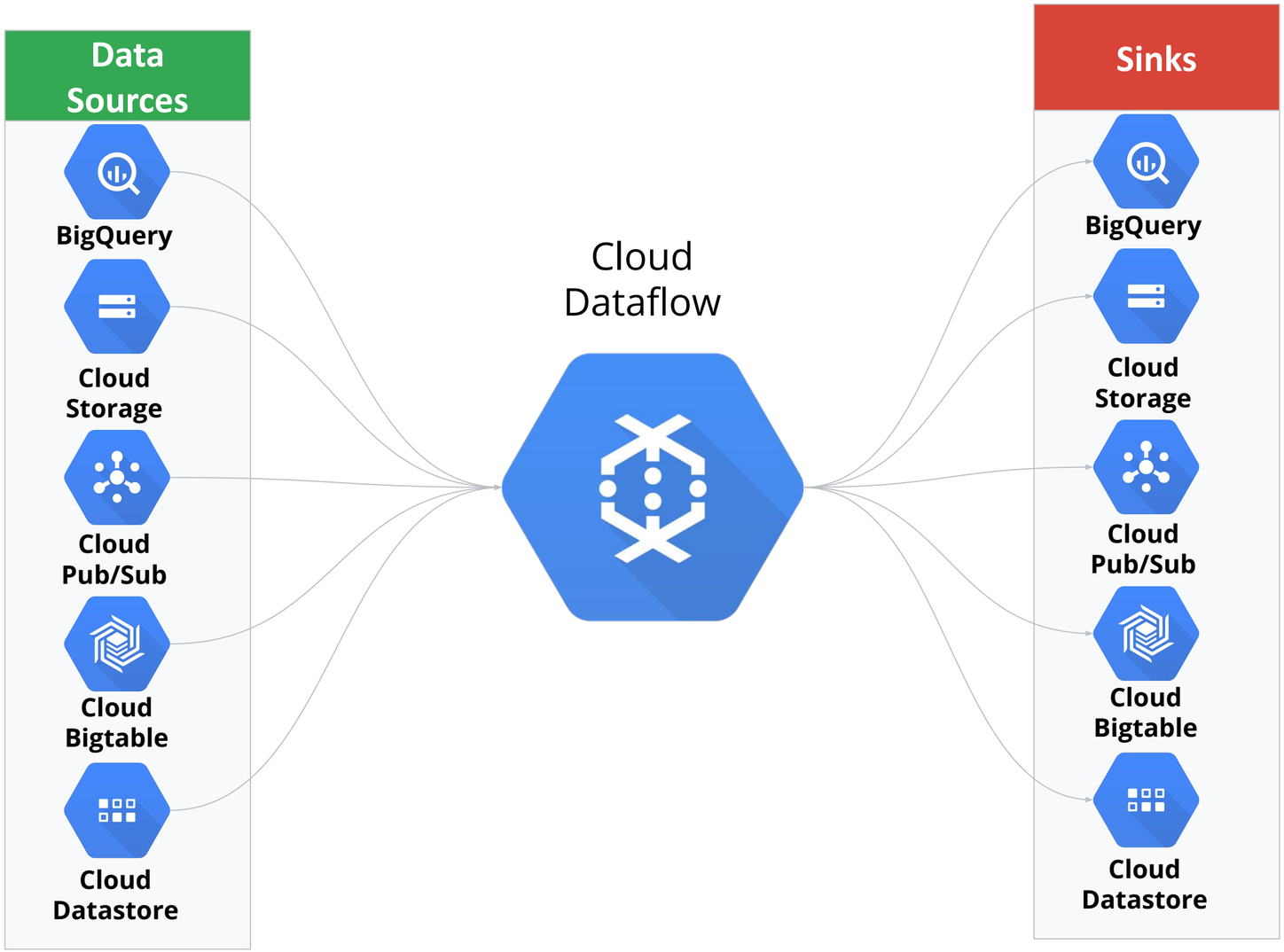
Open Source ensures Portability

- Pipelines written in Beam API are portable
- Runners include Dataflow, Flink, Samza and Spark

Auto-optimizations

- No more cluster management
- Submit a job and Dataflow auto-optimizes resources
- Makes pipelines faster and cost-effective

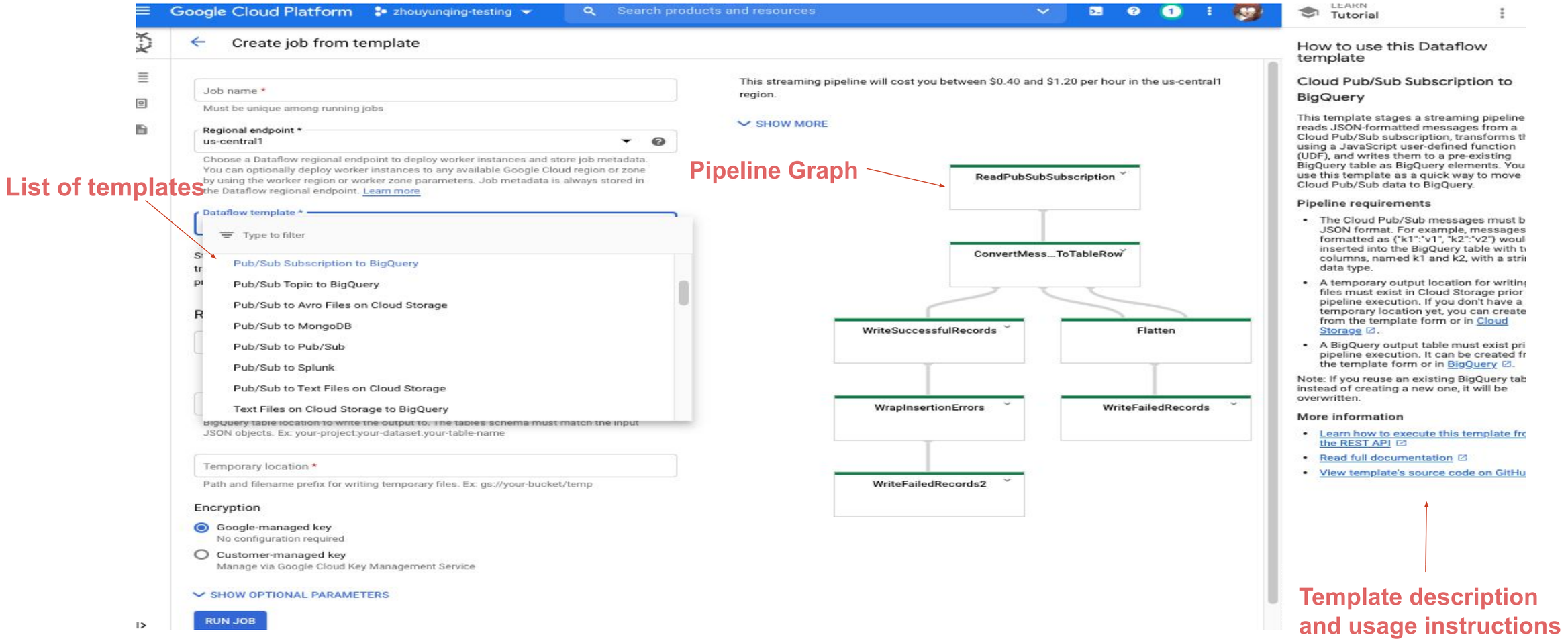
Data Sources and Sinks for Dataflow



Exam Tips:

- *Dataflow does NOT store data! There is always a Source and a Sink*

Dataflow: Google Provides Templates for different use-cases

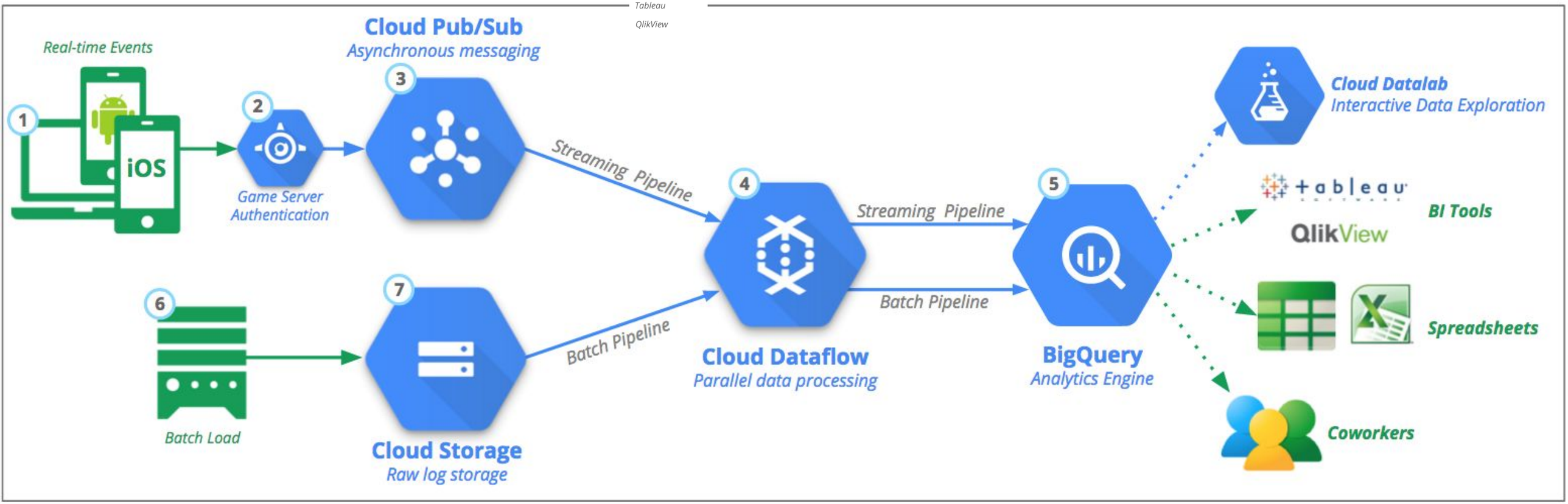


List of templates

Pipeline Graph

Template description and usage instructions

Most common ETL architecture



Diagnostic Question Discussion

Your company has successfully migrated to the cloud and wants to analyze their data stream to optimize operations. They do not have any existing code for this analysis, so they are exploring all their options. These options include a mix of batch and stream processing, as they are running some hourly jobs and live- processing some data as it comes in.

Which technology should they use for this?

A. Google Cloud Dataproc

B. Google Cloud Dataflow

C. GKE with Bigtable

D. Google Compute Engine with BigQuery

Diagnostic Question Discussion

Your company has successfully migrated to the cloud and wants to analyze their data stream to optimize operations. They do not have any existing code for this analysis, so they are exploring all their options. These options include a mix of batch and stream processing, as they are running some hourly jobs and live- processing some data as it comes in.

Which technology should they use for this?

A. Google Cloud Dataproc

B. Google Cloud Dataflow

C. GKE with Bigtable

D. Google Compute Engine with BigQuery

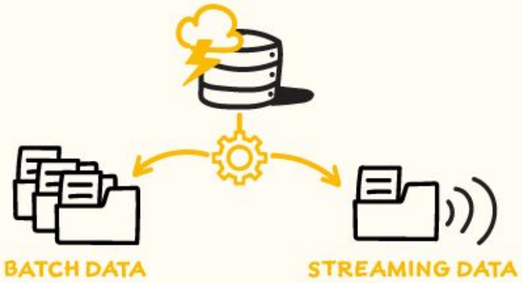
Dataflow

#GCPSketchnote

@PVERGADIA THECLOUDGIRL.DEV

What is Dataflow?

SERVERLESS, DATA PROCESSING SERVICE FOR BOTH STREAMING AND BATCH DATA



CREATE DATA PROCESSING PIPELINE

Apache Beam (SDK)

DEPLOY & EXECUTE

Dataflow

AUTOMATE
Operations & Management

Data Pipeline → Dataflow Job

How does it WORK?

DATA PROCESSING PIPELINE

STEP 1
Read Data Source

Files, BigQuery, Bigtable, Pub/Sub, Custom Source

STEP 2
Transform Data

PCollection 1 → PCollection 2 → PCollection 3

STEP 3
Write Data Sink

Files, BigQuery, Bigtable, Pub/Sub, Custom Source

Dataflow

WORKER VIRTUAL MACHINES EXECUTE THE DATA PROCESSING

Worker Virtual Machines

Autoscale up or down dynamically as needed

Dataflow Streaming Engine

Separate Compute | Separate Storage

BATCH PIPELINES SCALE SEAMLESSLY, WITHOUT ANY TUNING REQUIRED.

Dataflow → FEATURE Service-based Dataflow Shuffle → Data Grouping & Joining → Scale to 100s of TBs

HOW TO USE Dataflow

1 DATAFLOW SQL VIA BIGQUERY

- Use SQL from BigQuery web UI
- Read from pub/sub, cloud storage or BigQuery
- Write to BigQuery

2 DATAFLOW TEMPLATES

- Share pipeline with teams
- Easy & repeatable Pre built templates

3 AI PLATFORM NOTEBOOKS

- Use latest data science and machine learning frameworks

SECURITY for Dataflow Pipelines

EXACTLY-ONCE PROCESSING

0 Avoid risk of data loss

1 Turn off public IPs

2 VPC controls

3 Customer managed encryption key (CMEK)

Dataflow PRICING

DATAFLOW
Per second billing

DATAFLOW JOB
Batch or Streaming
(FlexRS) flexible resource scheduling = low cost

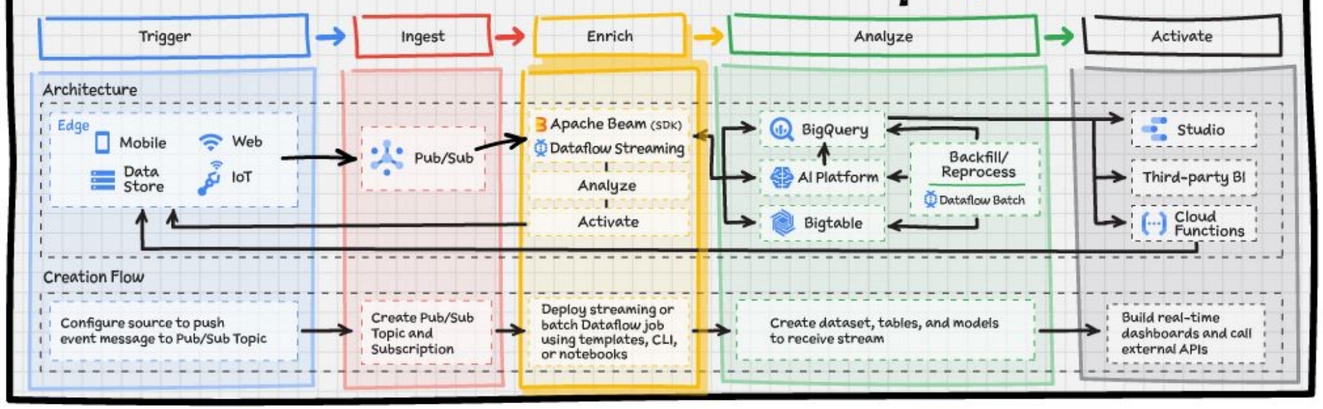
AUTOSCALING
Worker(s)
Better price to performance

CUSTOMIZABLE
vCPU + Memory + Storage
Each billed on per second basis

Dataflow Use case example

Streaming analytics

Real-time AI



Optional materials 1

[READING]

- get a feeling of the differences between PD snapshots, images and machine images (important from exam perspective:
- What is GCP metadata server?
- Sole-tenant nodes
- How stateful workloads are different from stateless workloads
- Image management best practices | Compute Engine Documentation | Google Cloud

[VIDEOS]

- Networking 102 (Cloud Routing and VPC Peering): Cloud OnAir: CE Chat: Google Cloud Networking 102 - Cloud Routing and VPC Peering
- GCE Managed Instance Groups: Using managed instance groups
- Shared VPC: Level Up From Zero Episode 4: Shared VPC
- BeyondCorp overview: BeyondCorp Enterprise in a minute
- Cloud security basics: Top 3 access risks in Cloud Security

Make sure to…

Enjoy the journey as much as the destination!