Google Cloud Platform

# Google Compute Engine and Networking

Google Cloud Platform Fundamentals

V2.1



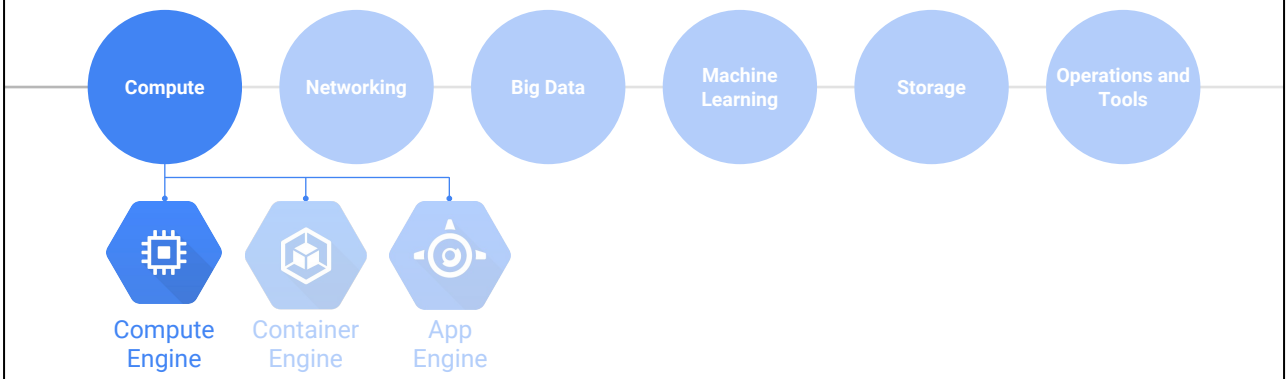


*Timing: Approximately 30 minutes*

# Agenda

1. Google Compute Engine Overview
2. Google Cloud Networking
3. Operations and Tools
4. Quiz & Lab

# Google Cloud Platform








Notes:
Compute Engine is one of several Google Cloud Platform compute options for running your applications.

# Google Compute Engine (1 of 3)

- Run large-scale workloads on virtual machines hosted on Google's infrastructure
- Robust networking features
  - Default, custom networks
  - Firewall rules
  - Regional HTTP(s) load balancing
  - Network load balancing
  - Subnetworks





Notes:
Google Compute Engine lets you create and run virtual machines on Google infrastructure. Compute Engine offers scale, performance, and value that allows you to easily launch large compute clusters on Google's infrastructure. There are no upfront investments and you can run thousands of virtual CPUs on a system that has been designed to be fast, and to offer strong consistency of performance.

An instance is a virtual machine (VM) hosted on Google's infrastructure. You can create an instance by using the Google Cloud Platform Console or the gcloud command-line tool. A Compute Engine instances can run Linux and Windows Server images provided by Google, or any customized versions of these images. You can also build and run images of other operating systems. You can choose the machine properties of your instances, such as the number of virtual CPUs and the amount of memory, by using a set of predefined machine types or by creating your own custom machine types.

Your networks connect your instances to each other and to the Internet. You can segment your networks, use firewall rules to restrict access to instances, and create static routes to forward traffic to specific destinations.

Scale your applications on Google Compute Engine from zero to full-throttle

with Google Cloud Load Balancing, with no pre-warming needed. Distribute your load-balanced compute resources in single or multiple regions, close to your users and to meet your high availability requirements. Cloud Load Balancing can put your resources behind a single anycast IP and scale your resources up or down with intelligent autoscaling. Cloud Load Balancing comes in a variety of flavors and is integrated with Google Cloud CDN for optimal application and content delivery.

Subnetworks segments your Cloud network IP space into subnetworks. Subnetwork prefixes can be automatically allocated, or you can create a custom topology. For more information on subnetworks, see: https://cloud.google.com/compute/docs/subnetworks.

# Google Compute Engine (2 of 3)

- High CPU, high memory, standard and shared-core machine types
- Persistent disks
  - Standard, SSD, local SSD
  - Snapshots
- Resize disks, migrate instances with no downtime
- Instance metadata and startup scripts

# Google Compute Engine (3 of 3)

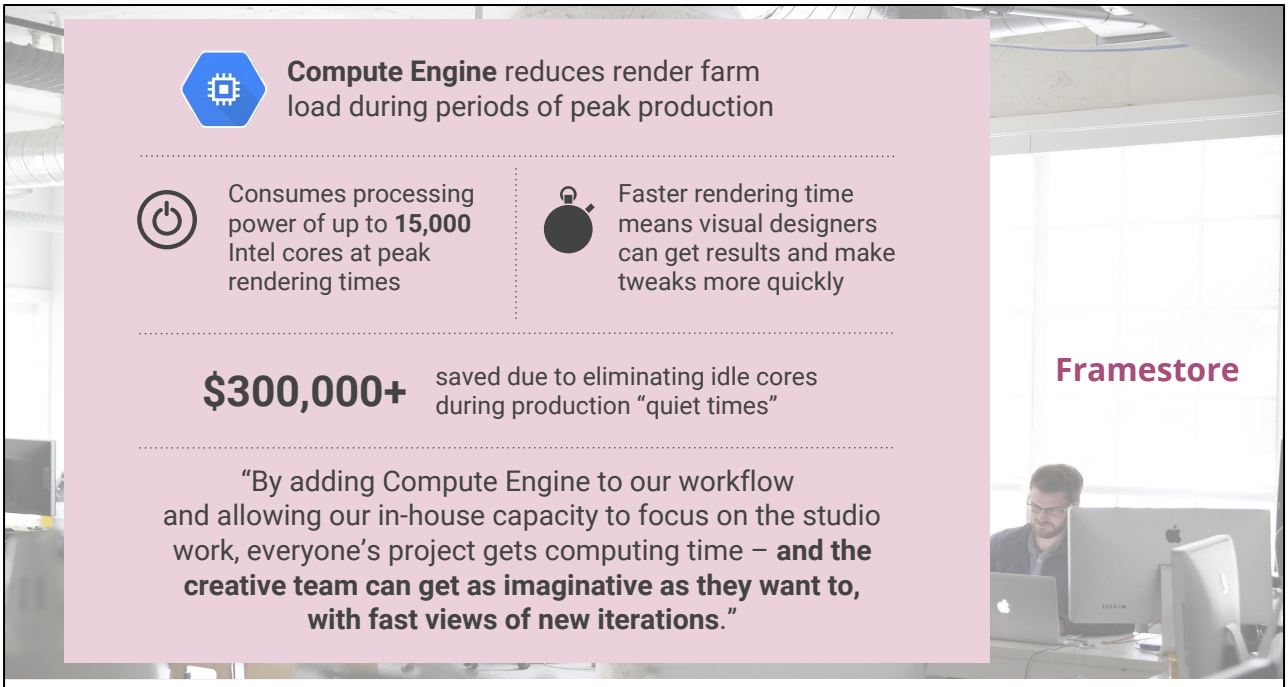
- Advanced APIs for auto-scaling and instance group management
- Innovative pricing
  - **Per-minute** billing, sustained use discounts
  - Preemptible instances
  - High throughput to storage at no extra cost
  - Custom machine types - Only pay for the hardware you need





Notes:
For more info on preemptible instances, go to:
https://cloud.google.com/compute/docs/instances/preemptible.

For more information on using custom machine types, see:
https://cloud.google.com/compute/docs/instances/creating-instance-with-custom-machine-type.




Notes:
Read more about the Framestore story at:
https://cloudplatform.googleblog.com/2014/11/framestore-frees-up-designers-to-create-unforgettable-visual-effects-with-help-from-google-compute-engine.html

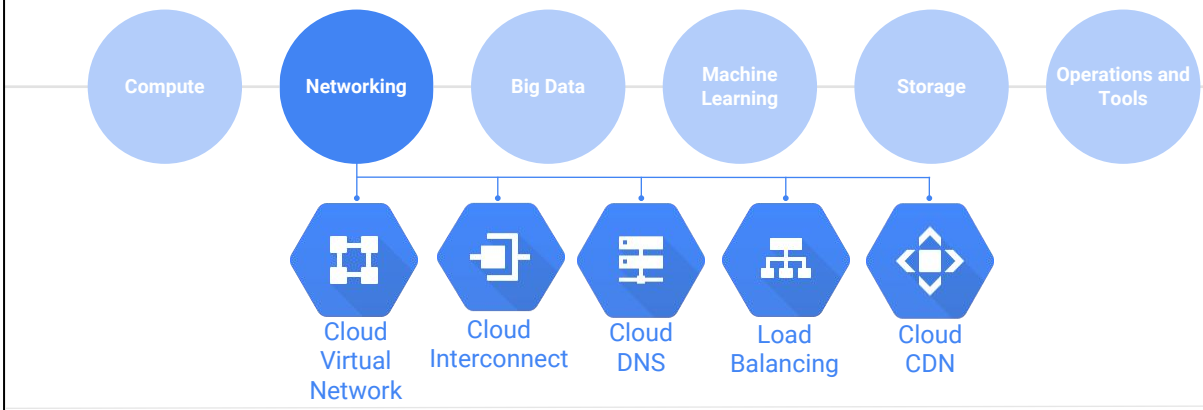
# Comparing Compute Options

| | Compute Engine | Container Engine | App Engine Standard | App Engine Flexible |
|---|---|---|---|---|
| *Language support* | Any | Any | Java, Python, Go & PHP | Any |
| *Service model* | IaaS | Hybrid | PaaS | PaaS |
| *Primary use case* | General computing workloads | Container-based workloads | Web and mobile applications | Web and mobile applications, container-based workloads |

# Agenda

1. Google Compute Engine Overview
2. Google Cloud Networking
3. Operations and Tools
4. Quiz & Lab

# Google Cloud Platform








Notes:
Google Cloud Platform has a number of networking services to support your cloud applications, including: Google Cloud Virtual Network, Google Cloud Interconnect, Google Cloud DNS, Google Cloud Load Balancing, and Google Cloud CDN.

# Cloud Virtual Network (1 of 2)

- Managed networking functionality for Cloud Platform resources
- Provision Google Cloud Platform resources, connect them to each other and isolate them from one another in a Virtual Private Cloud (VPC)





Notes:
**Cloud Virtual Network features**

*Network*

- Create a legacy network with a single global IP range, or a subnetwork that is regional. A subnet network can be of auto mode or custom mode. Subnets size in custom networks can dynamically increase its size by expanding CIDR range, and can move from one availability zone to another (from A to B) within a region (such as us-west1), without any impact to configured VMs.

# Cloud Virtual Network (2 of 2)

- Cloud Virtual Network includes:
  - Fine-grained networking policies
  - Granular IP address range selection
  - Routes
  - Firewalls
  - Virtual Private Network (VPN)
  - Cloud Router





Notes:
**Cloud Virtual Network features**

*Routes*
- Forward traffic from one instance to another instance within the same network, even across subnetworks, without requiring an external IP addresses.

*Firewalls*
- Segment your networks with a global distributed firewall to restrict access to instances.

*VPN*
- Securely connect your existing network to Compute Engine network over IPsec - Encrypts traffic over the Internet

*Cloud Router*
- Enable dynamic Border Gateway Protocol (BGP) route updates between your Compute Engine network and your non-Google network with our virtual router.

# Google Cloud Interconnect

| *Carrier Interconnect* | *Direct Peering* | *CDN Interconnect* |
|---|---|---|
| Enterprise-grade connections provided by carrier service providers | Connect your business directly to Google | Allows select CDN providers to establish direct interconnect links with Google's edge network at various locations |

Google Cloud Interconnect





Notes:
Google Cloud Interconnect allows Google Cloud Platform customers to connect to Google via enterprise-grade connections with higher availability and/or lower latency than their existing Internet connections. Connections are offered by Carrier Interconnect service provider partners, and may offer higher SLAs than standard Internet connections. Google also supports direct connections to its network through direct peering. Customers who cannot meet Google at its peering locations, or do not meet peering requirements, may benefit from Carrier Interconnect.

# Google Cloud DNS

- Highly available and scalable DNS
  - Translates domain names into IP addresses
- Create managed zones, then add, edit, delete DNS records
  - Programmatically manage zones and records using RESTful API or command-line interface





Notes:
Google Cloud DNS is a scalable, reliable and managed authoritative Domain Name System (DNS) service running on the same infrastructure as Google. It has low latency, high availability and is a cost-effective way to make your applications and services available to your users. Cloud DNS translates requests for domain names like www.google.com into IP addresses like 74.125.29.101. Cloud DNS is programmable. You can easily publish and manage millions of DNS zones and records using our simple user interface, command-line interface or API.

Cloud DNS uses our global network of Anycast name servers to serve your DNS zones from redundant locations around the world, providing high availability and lower latency for your users.

# Google Cloud Load Balancing (1 of 2)

- HTTP(s) load balancing
  - Balance HTTP-based traffic across multiple Compute Engine regions
  - Global, external IP address routes traffic
  - Scalable, requires no pre-warming and provides resilience, fault tolerance





Notes:
Scale your applications on Google Compute Engine from zero to full-throttle with Google Cloud Load Balancing, with no pre-warming needed. Distribute your load-balanced compute resources in single or multiple regions, close to your users and to meet your high availability requirements. Cloud Load Balancing can put your resources behind a single anycast IP and scale your resources up or down with intelligent Autoscaling. Cloud Load Balancing comes in a variety of flavors and is integrated with Google Cloud CDN for optimal application and content delivery.

With Cloud Load Balancing, a single anycast IP front-ends all your backend instances in regions around the world. It provides cross-region load balancing including automatic multi-region failover which gently moves traffic in fractions if backends become unhealthy. In contrast to DNS-based Global Load Balancing solutions, Cloud Load Balancing reacts instantaneously to changes in users, traffic, network, backend health and other related conditions.

Cloud Load Balancing is a fully distributed, software-defined, managed service for all your traffic. It is not an instance or device based solution, so you won't be locked into physical load balancing infrastructure or face the HA, scale and management challenges inherent in instance based LBs. You can apply Cloud Load Balancing to all of your traffic: HTTP(S), TCP/SSL, and UDP. You can also

terminate your SSL traffic with HTTPS Load Balancing and SSL proxy.

Cloud Load Balancing can scale as your users and traffic grow, including easily handling huge, unexpected and instantaneous spikes by diverting traffic to other regions in the world that can take traffic. Autoscaling does not require pre-warming, you can scale from zero to full throttle in matter of seconds.

# Google Cloud Load Balancing (2 of 2)

- TCP/SSL and UDP (network) load balancing
  - Spread TCP/SSL and UDP traffic over pool of instances within a Compute Engine region
  - Ensures only healthy instances handle traffic
  - Scalable, requires no pre-warming





Notes:
Network load balancing allows you to balance load of your systems based on incoming IP protocol data, such as address, port, and protocol type.

Network load balancing uses forwarding rules that point to target pools, which list the instances available for load balancing and define which type of health check that should be performed on these instances.

Network load balancing offers some load balancing options that are not available with HTTP(S) load balancing. For example, you can load balance additional TCP/UDP-based protocols such as SMTP traffic. If your application is interested in TCP-connection-related characteristics, network load balancing allows your app to inspect the packets, which you cannot do with HTTP(S) load balancing.

You can also use network load balancing to handle HTTPS traffic. You will need to:

- Create a firewall rule that allows traffic on port 443, and apply it to the balanced instances.
- Handle the encryption/decryption on your VM instances.

If you need cross-regional load balancing, use HTTP(S) load balancing.

# Google Cloud CDN (Content Delivery Network)

- Use Google's globally distributed edge caches to cache HTTP(S) load balanced content far closer to your users than your instances
  - Faster delivery of content to users while reducing costs
- Cloud CDN uses caches at network locations to store responses generated by instances





Notes:
Google Cloud CDN leverages Google's globally distributed edge caches to accelerate content delivery for websites and applications served out of Google Compute Engine. Cloud CDN lowers network latency, offloads origins, and reduces serving costs. Once you've set up HTTP(S) Load Balancing, simply enable Cloud CDN with a single checkbox.

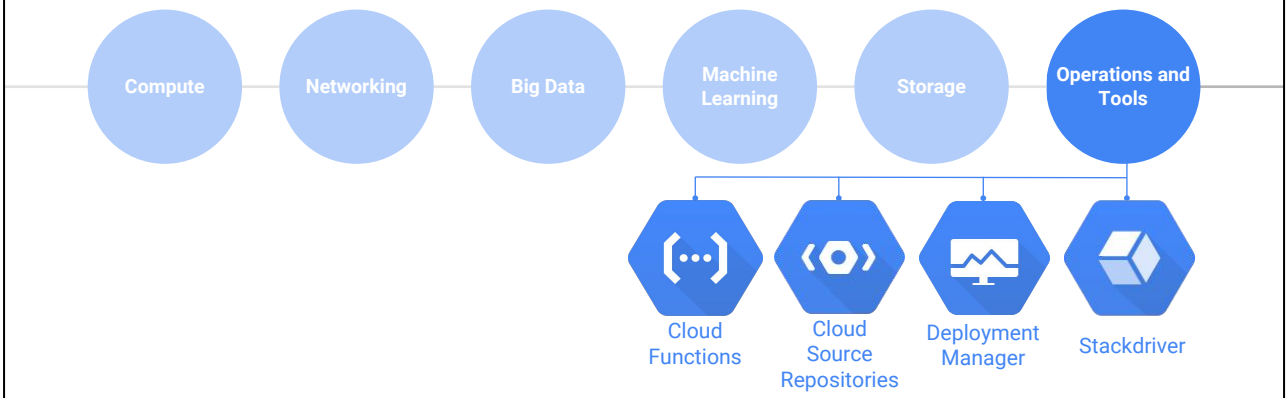
The Cloud CDN content delivery network works with Google Compute Engine HTTP(S) load balancing to deliver content to your users. Responses from your virtual machine (VM) instances flow through an HTTP(S) load balancer before being delivered by Cloud CDN. When a user requests content from your site, that request passes through network locations at the edges of Google's network, usually far closer to the user than your instances. Cloud CDN uses caches at these network locations to store responses generated by your instances.

For more information on Google Cloud CDN, see:
https://cloud.google.com/cdn/docs/overview.

# Agenda

1. Google Compute Engine Overview
2. Google Cloud Networking
3. Operations and Tools
4. Quiz & Lab

# Google Cloud Platform








Notes:
Google Cloud Platform has a number of management and developer tools to support your cloud applications, including: Google Cloud Functions, Google Cloud Source Repositories, Deployment Manager, and Stackdriver. Many other tools are available as well. For a complete list of management and development tools, see: https://cloud.google.com/products/.

# Google Stackdriver Beta (1 of 2)

- Integrated monitoring, logging, diagnostics
- Works across Google Cloud Platform, Amazon Web Services
- Open source agents, integration
- Powerful data, analytics tools
- Collaborations with PagerDuty, BMC, Splunk, others





Notes:
Google Stackdriver provides powerful monitoring, logging, and diagnostics. It equips you with insight into the health, performance, and availability of cloud-powered applications, enabling you to find and fix issues faster. It is natively integrated with Google Cloud Platform, Amazon Web Services, and popular open source packages. Stackdriver provides a wide variety of metrics, dashboards, alerting, log management, reporting, and tracing capabilities.

Stackdriver gives you access to logs, metrics, traces, and other signals from your infrastructure platform(s), virtual machines, containers, middleware, and application tier, so that you can track issues all the way from your end user to your backend services and infrastructure. Native support for distributed systems, auto-scaling, and ephemeral resources means that your monitoring works seamlessly with your modern architecture.

# Google Stackdriver Beta (2 of 2)

## Monitoring

Platform, system, and application metrics
Uptime/health checks
Dashboards and alerts

## Logging

Platform, system, and application logs
Log search/view/filter
Log-based metrics

## Trace

Latency reporting and sampling
Per-URL latency and statistics

## Error Reporting

Error notifications
Error dashboard

## Debugger

Debug applications





Notes:
The core components of Google Stackdriver include: monitoring, logging, trace, error reporting, and debugging.

Stackdriver Monitoring provides endpoint checks to web applications and other internet-accessible services running on your cloud environment. You can configure uptime checks associated with URLs, groups, or resources, such as instances and load balancers.

Stackdriver Logging provides you with the ability to filter, search, and view logs from your cloud and open source application services. Logging also allows you to define metrics based on log contents that are incorporated into dashboards and alerts, and it enables you to export logs to BigQuery, Google Cloud Storage, and Pub/Sub.

Stackdriver Trace provides latency sampling and reporting for Google App Engine, including per-URL statistics and latency distributions.

Stackdriver Error Reporting analyzes and aggregates the errors in your cloud applications, notifies you when new errors are detected, and alerts allow you to create alerting policies to notify you when metrics, health check results, and uptime check results meet specified criteria. Error reporting is Integrated with

a wide variety of notification channels, including Slack, Campfire, Hipchat, and Pagerduty. Dashboards provide you with default dashboards for cloud and open source application services and allow you to define custom dashboards with powerful visualization tools to suit your needs.

Stackdriver Debugger connects your application's production data to your source code by inspecting the state of your application at any code location in production without stopping or slowing down your requests. The debugger makes it easier to view the application state without adding logging statements.

# Google Cloud Deployment Manager

- Infrastructure management service
- Create a `.yaml` template describing your environment and use Deployment Manager to create resources
- Provides repeatable deployments





Notes:
Deployment Manager is an infrastructure management service that automates the creation and management of your Google Cloud Platform resources for you. Using Deployment Manager, create flexible and declarative templates that deploy a variety of Cloud Platform services, such as Google Cloud Storage, Google Compute Engine, Google Cloud SQL, and leave it to Deployment Manager to instantiate and manage the Cloud Platform resources defined in your templates as deployments.

# Google Cloud Source Repositories Beta

- Fully-featured Git repositories hosted on Google Cloud Platform
- Supports collaborative development of cloud apps
- Includes integration with Stackdriver debugger





Notes:
Google Cloud Source Repositories provides Git version control to support collaborative development of any application or service, including those that run on Google App Engine and Google Compute Engine. If you are using the Stackdriver Debugger, you can use Cloud Source Repositories and related tools to view debugging information alongside your code during application runtime. Cloud Source Repositories also provides a source viewer that you can use to browse and view repository files from within the Cloud Console.

With Cloud Source Repositories, you can have any number of private Git repositories, allowing you to organize the code associated with your cloud project in whatever way works best for you. Google Cloud diagnostics tools like the Debugger and Error Reporting can use the code from your git repositories to let you track down issues to specific errors in your deployed code without slowing down your users. If you've already got your code in GitHub or BitBucket repositories, you can bring that into your cloud project and use it just like any other repository, including browsing and diagnostics.

# Google Cloud Functions Alpha

- Create single-purpose functions that respond to events without a server or runtime
  - Event examples: New instance created, file added to Cloud Storage
- Written in Javascript, execute in managed Node.js environment on Google Cloud Platform





Notes:
Google Cloud Functions is a lightweight, event-based, asynchronous compute solution that allows you to create small, single-purpose functions that respond to cloud events without the need to manage a server or a runtime environment.

Cloud Functions are written in Javascript and execute in a managed Node.js environment on Google Cloud Platform. Events from Google Cloud Storage and Google Cloud Pub/Sub can trigger Cloud Functions asynchronously, or you can use HTTP invocation for synchronous execution.

Cloud Events are things that happen in your cloud environment. These might be things like changes to data in a database, files added to a storage system, or a new virtual machine instance being created.
Events occur whether or not you choose to respond to them. Creating a response to an event is done with a Trigger. A trigger is a declaration that you are interested in a certain event or set of events. You create triggers to capture events and act on them.

# Agenda

1. Google Compute Engine Overview
2. Google Cloud Networking
3. Operations and Tools
4. Quiz & Lab

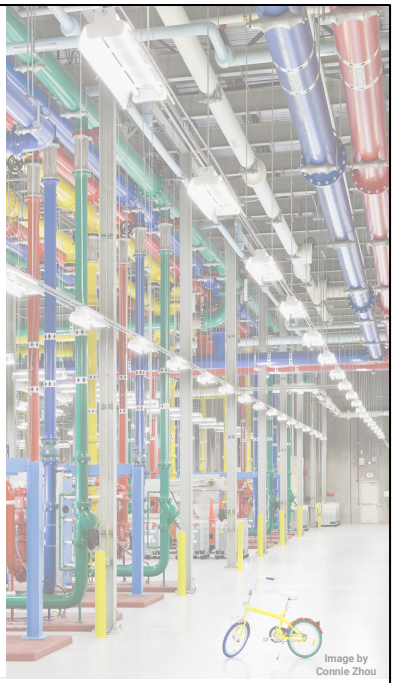
# Quiz

1. Name 3 robust networking services available to your applications on Google Cloud Platform.
2. Name 3 Compute Engine pricing innovations.
3. *True or False*: Google Cloud Load Balancing allows you to Balance HTTP-based traffic across multiple Compute Engine *regions*.

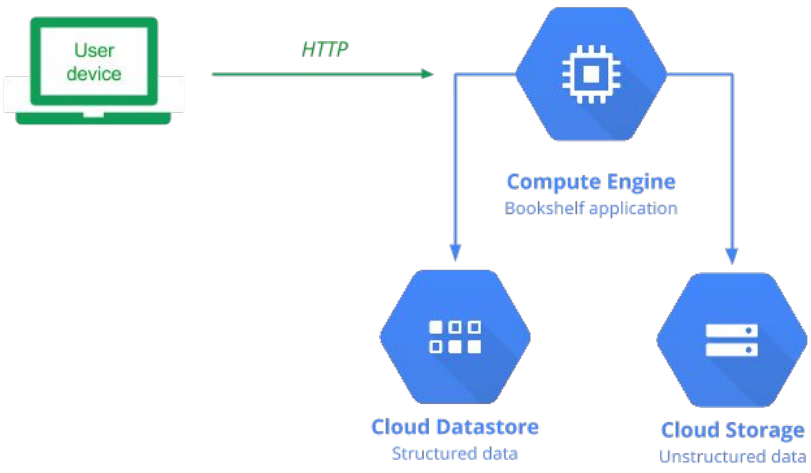
# Lab (1 of 2)

Deploy the Bookshelf application to Compute Engine.

1. Create a Google Compute Engine instance
2. Deploy the Bookshelf application using a startup script
3. Add a firewall rule to allow HTTP traffic
4. Test the Bookshelf application in your browser



# Lab (2 of 2)








Notes:
In this lab, you create a Google Compute Engine instance and deploy the Bookshelf application to it using a Linux shell startup script. You continue to use existing storage services you used earlier in the course, including Google Cloud Datastore, and Google Cloud Storage.

# Resources

- Google Compute Engine
  https://cloud.google.com/compute/docs/
- Google Cloud CDN
  https://cloud.google.com/cdn/docs/
- Google Cloud Stackdriver
  https://cloud.google.com/stackdriver/docs/
- Google Cloud Deployment Manager
  https://cloud.google.com/deployment-manager/docs/
- Google Cloud Source Repositories
  https://cloud.google.com/source-repositories/docs/

# Quiz Answers

1. Name 3 robust networking services available to your applications on Google Cloud Platform.

   *Answer*: Cloud Virtual Network, Cloud Interconnect, Cloud DNS, Cloud Load Balancing, and Cloud CDN.

2. Name 3 Compute Engine pricing innovations.

   *Answer*: Per-minute billing, custom machine types, preemptible instances.

3. *True*: Google Cloud Load Balancing allows you to Balance HTTP-based traffic across multiple Compute Engine *regions*.


cloud.google.com