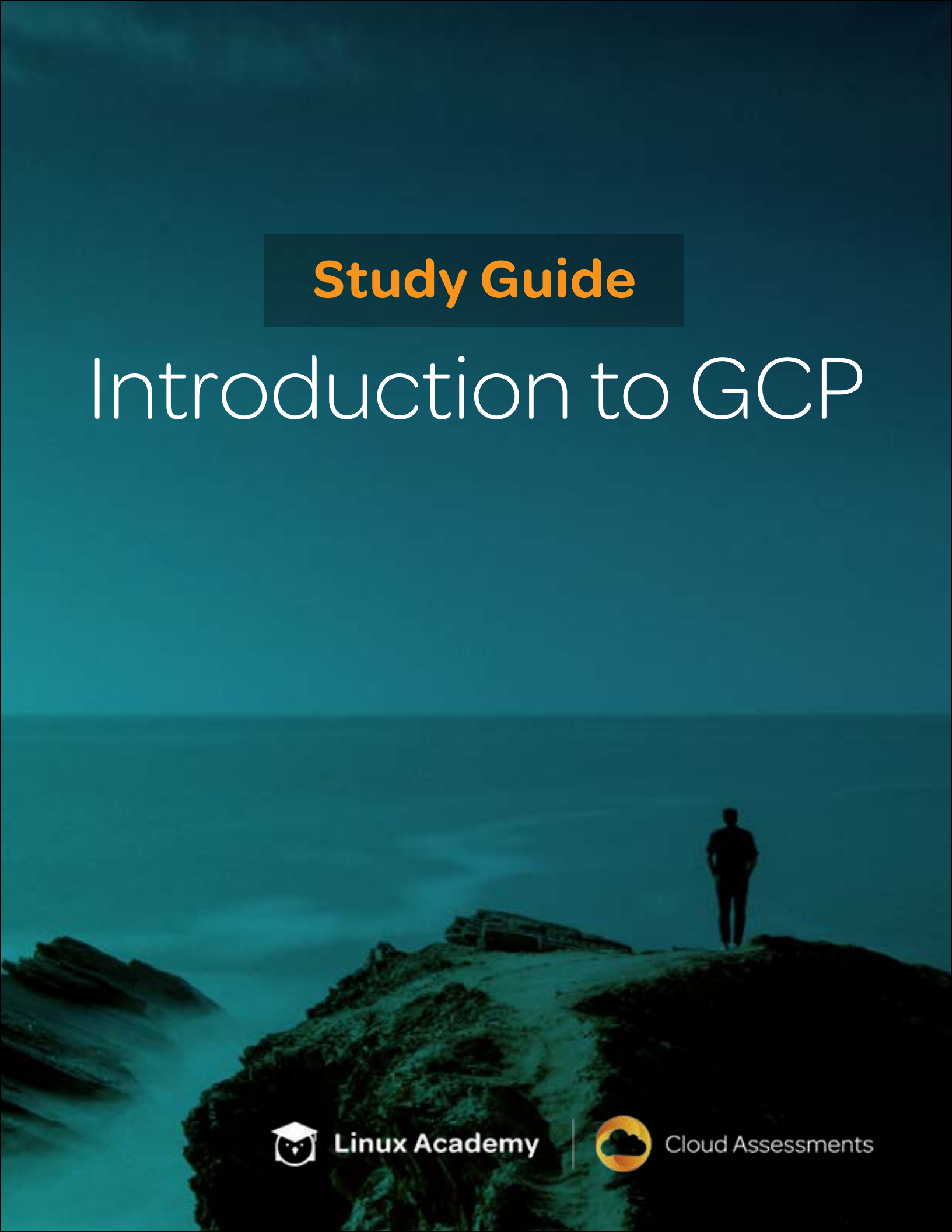
Study Guide
Introduction to GCP
Linux Academy
Cloud Assessments

# Contents

# Why Choose the Google Cloud Platform?

What is Google Cloud Platform?

- Suite of cloud computing services
- Google's worldwide collection of data centers
- Hosts IaaS, PaaS, and SaaS use cases
- At a basic level, it hosts and manages your computer infrastructure so you don't have to

Like most other cloud platforms:

- "Pay-as-you-go" basis, use only what you need
    - Also known as 'on demand' computing
- Convert capital expenses into operating expenses
- Focus on rapid innovation
- Productivity enhanced due to no software installed
- "Vertically-integrated" stacks enhance functionality, performance, reliability, and security

What makes GCP stand out?

- Access the same infrastructure Google uses for their own services
- Customer-friendly pricing (https://cloud.google.com/pricing)
    - Compute Engine instances billed per minute, not hour
    - Automatic sustained use discounts; no upfront commitment required
    - Archive coldline storage for extremely cheap
- Private Global Fiber Network
    - Blazing fast
    - Data moving between data centers never leaves the private network

- Live migration of virtual machines
- Better performance
    - Website benchmarks showed as much as 50% improvement in loading speed
    - Instantly scalable; VMs can auto-scale without shutting machine down
    - Access to GPUs for high-end scientific computing and machine learning
- Industry leading security
    - Over 500 full time security professionals
    - All data encrypted in transit and at rest
    - Expansive audit compliance list
    - SSAE16, ISO 27
    - 017, ISO 27018, PCI, and HIPAA
- Access to innovative resources not available anywhere else
    - Big Data, Machine Learning
    - APIs (Video Intelligence, Maps)
    - “Build what’s next”

## Data Center Infrastructure

How is GCP built?

- Data centers
    - Google operates an extensive deployment of high-efficiency backend data centers used for computation and storage capacity.
- Backbone
    - Google has built a global, meshed backbone network to connect their data centers and to deliver traffic to their edge points of presence (POPs).

- Points of presence
    - 90+ edge points of presence in 33 countries connected via Google's backbone network.
- Edge caching
    - Google runs an edge caching platform on top of their network infrastructure with edge locations in virtually every country.

Cloud Regions and Zones

- Nine regions split into 27 zones
- Regions
    - Regions are specific geographical locations where you can run your resources
    - Collections of zones
    - Regional resources are available to resources in any zone in the region
    - Frequently expanding
- Zones
    - Isolated physical locations within a region
    - Zonal resources are only available in that zone
    - Machines in different zones have no single point of failure
- For example: an effective disaster recovery plan would have assets deployed across multiple zones, or even different regions.

Commitment to environmental responsibility

- 100% renewable energy
- Extremely efficient infrastructure
- Optimized to run within Google's environment
- Carbon neutral since 2007

- Hosting company resources on GCP is more environmentally-friendly than hosting your own server resources
- Learn more at https://www.google.com/about/datacenters

Network infrastructure from data centers to end users

- Three elements:
  - Core data centers
  - Edge Points of Presence (PoPs)
  - Edge nodes

Points of Presence

- 90+ locations worldwide
- Brings Google traffic closer to users worldwide, thereby reducing their costs and providing users with a better experience.
- Connects to the private meshed network backbone that connects Edge PoPs to data centers and bridges to the public Internet

Edge Nodes

- Tier of Google's infrastructure closest to end users
- Internet service providers (ISPs) deploy Google-supplied servers inside their own network
- "Static content that is very popular with the local host's user base, including YouTube and Google Play, is temporarily cached on edge nodes. Google's traffic management systems direct user requests to an edge node that will provide the best experience."
- Pulling popular content from edge cache is substantially faster than pulling everything from data centers

End result

- Fast, redundant, worldwide presence that provides fast and reliable access to your resources no matter where in the world you are
- All of this while being 100% carbon neutral

# Certification Overview

Certification track has drastically changed in the past year

- Any certification information before December 2016 is no longer valid
- There are now only three certification exams, two of which relate to Google Cloud
    - Google Cloud Architect
    - Google Cloud Data Engineer
    - G Suite Administrator
- The first two are considered professional-level exams and most closely correlate with AWS's professional level exams. In other words, there is no intermediate exam level.

Cramming for exam purposely made difficult

- Per Google's cert team:
    - There are no formal requirements to take the Google Cloud Architect Certification exam. However, please note that hands-on experience performing the job tasks in the guide and using Google Cloud Platform technology is the best preparation for the exam. Google's certifications are intended to identify people who have demonstrated skills required for a job role. That's why we do not offer exam prep materials, as the exams are not intended to assess someone's mastery of training.

Training tracks vs. certification exams

- Old exam info is now rolled into Google's training track site. Material is largely the same, however there is no certification test to take for it.
- https://cloud.google.com/training
- This course will follow the Google Cloud Platform Fundamentals: Core Infrastructure objectives, and will serve as a foundation for more advanced courses.
- Ironically, Google hasn't finished defining their own advanced-level track.

Where does this leave us?

- Learn by doing: The best way to prepare for the Cloud Architect exam is to use GCP, become intimately familiar with all of its services, and know what service serves what purpose.

- This course will act as the foundation to build upon for more advanced concepts.
- Our courses will stress working hands on and will give exercises to use it to achieve business objectives.
- We will explain both the hands-on and high-level concepts necessary to pass the exam as we ramp up our courses in skill level.

# Getting Started with Google Cloud Platform

## Organizations and Projects.

### GCP Organization Hierarchy

Just like any operating system or organization with multiple people requiring different types of access, some form of organization and access control is a necessity.

GCP's implementation is called the Cloud Resource Hierarchy. To use an analogy, the Cloud Resource Hierarchy is similar to the file system found in traditional operating systems as a way of organizing and managing entities in a hierarchical format. Each resource has exactly one parent. This hierarchical organization of resources enables you to set access control policies and configuration settings on a parent resource, and the policies and IAM settings are inherited by the child resources.

The purpose of the Cloud Resource Hierarchy is two-fold:

- Provide a hierarchy of ownership that binds the lifecycle of a resource to its immediate parent in the hierarchy
- Provide attach points and inheritance for access control and organization policies

### Organization

The Organization resource is the root node of the Cloud Resource Hierarchy and all resources that belong to an organization are grouped under the organization node. This provides central visibility and control over every resource that belongs to an organization.

### High-level Overview of This Hierarchy

- Highest level is an Organization, which are divided into Projects, which are divided further into Resources
- All Google Cloud Platform services are associated with a project that is used to:
    - Track resource and quota usage
    - Enable billing
    - Manage permissions and credentials
    - Enable services and APIs

- Projects use three identifying attributes:
    - Project Name
    - Project Number
    - Project ID
        - Also known as Application ID

## Identity and Access Management (IAM)

Why is this important?

- Any organization needs to determine who can access what resources. This is true in both inside and outside the technology realm.
- Cloud IAM lets you adopt the security principle of least privilege, so you grant only the necessary access to your resources and prevent unwanted access to other resources. IAM allows you to meet compliance clauses around the separation of duty.
- A more basic way of stating this is that IAM lets you manage access control by defining who (members) has what access (role) for which resource.

Let's break down those categories:

- How is access managed?
    - Members (who), are granted permissions and roles (what)
- What are members?
    - Google account
        - A Google account represents a developer, an administrator, or any other person who interacts with Google Cloud Platform. An email address that is associated with a Google account, such as a gmail.com address, can be an identity. New users can sign up for a Google account by going to the Google account signup page.

- Service account
    - A service account is a special type of Google account that belongs to your application or a virtual machine (VM), instead of to an individual end user. Your application calls Google APIs assuming the identity of the service account, so that the users aren't directly involved.
    - Provide an identity for carrying out server-to-server interactions in a project
    - Used to authenticate from one service to another
    - Can be used with primitive and curated roles
    - Identified with an email address:
        - '@developer.gserviceaccount.com' '@developer.gserviceaccount.com'
    - A service account is an account that belongs to your application instead of to an individual end user. When you run code that is hosted on Cloud Platform, you specify the account that the code should run as. You can create as many service accounts as needed to represent the different logical components of your application. See the Google Cloud Platform Console Service Accounts documentation for more information.
- Google group
    - A Google group is a named collection of Google accounts and service accounts. Every group has a unique email address that is associated with the group. You can find the email address that is associated with a Google group by clicking "About" on the homepage of any Google group. For more information about Google groups, see the Google groups homepage.
    - Google groups are a convenient way to apply an access policy to a collection of users. You can grant and change access controls for a whole group at once instead of granting or changing access controls one at a time for individual users or service accounts. You can also easily add members to and remove members from a Google group instead of updating a Cloud IAM policy to add or remove users.
    - Note that Google groups don't have login credentials, and you cannot use Google groups to establish identity to make a request to access a resource.

- G Suite domain
    - A G Suite domain represents a virtual group of all the members in an organization. G Suite customers can associate their email accounts with an internet domain name. When you do this, each email account takes the form username@yourdomain.com. You can specify an identity by using any internet domain name that is associated with a G Suite account.
    - Like groups, domains cannot be used to establish identity, but they enable convenient permission management.
- Cloud Identity Domain
    - A Cloud Identity domain is like a G Suite domain, because it represents a virtual group of all members in an organization. However, Cloud Identity domain users don't have access to G Suite applications and features.

## Roles

- Primitive Roles
    - Owner
    - Editor
    - Viewer
    - Billing Administrator
    - A project can have multiple owners, editors, viewers, and billing administrators.
    - Primitive roles: The roles historically available in the Google Cloud Platform Console will continue to work. These are the Owner, Editor, and Viewer roles.
    - Predefined roles: Predefined roles are the new IAM roles that give more finer-grained access control than the primitive roles. For example, the predefined role Publisher provides access to only publish messages to a Pub/Sub topic.
    - Prior to Cloud IAM, you could only grant Owner, Editor, or Viewer roles. These roles give very broad access on a project and did not allow separation of duties. Cloud Platform services now offer additional roles that give finer-grained access control than the Owner, Editor, and Viewer roles. For example, Compute Engine offers roles such as Instance Admin and Network Admin, while App Engine offers roles such as App Admin and Service Admin. These predefined roles are available in addition to the legacy Owner, Editor, and Viewer roles.

When would I use primitive roles?

- Use primitive roles in the following scenarios:
    - When the Cloud Platform service does not provide a predefined role. See the predefined roles table for a list of all available predefined roles.
    - When you want to grant broader permissions for a project. This often happens when you're granting permissions in development or test environments.
    - When you need to allow a member to modify permissions for a project, you'll want to grant them the owner role because only owners have the permission to grant access to other users for for projects.
    - When you work in a small team where the team members don't need granular permissions.

IAM Roles - Predefined Roles

- Much more granular access, prevent unwanted access to other resources
    - Granted at resource level
- Example: App Engine Admin – Full access to only App Engine resources
- Multiple predefined roles can be given to individual users
- All current Predefined Roles - https://cloud.google.com/iam/docs/understanding-roles#predefined_roles

## IAM Policy

You can grant roles to users by creating a Cloud IAM policy, which is a collection of statements that define who has what type of access. A policy is attached to a resource and is used to enforce access control whenever that resource is accessed.

## IAM Policy Hierarchy

Cloud Platform resources are organized hierarchically, where the Organization node is the root node in the hierarchy, the projects are the children of the Organization, and the other resources are the children of projects. Each resource has exactly one parent.

You can set an IAM access control policy at any level in the resource hierarchy: The Organization level, the project level, or the resource level. Resources inherit the policies of the parent resource. If you set a policy at the organization level, it is automatically inherited by all its children projects,

and if you set a policy at the project level, it is inherited by all its children resources. The effective policy for a resource is the union of the policy set at that resource and the policy inherited from its parent. This policy inheritance is transitive; in other words, resources inherit policies from the project, which inherit policies from the organization. Therefore, the organization-level policies also apply at the resource level.

For example, consider a Pub/Sub resource that lives under the project example-test. If you grant the editor role to bob@gmail.com for example-test, and grant the publisher role to alice@gmail.com for topic_a, you effectively grant editor role to bob@gmail.com and publisher role to alice@gmail.com for topic_a.

The IAM policy hierarchy follows the same path as the Cloud Platform resource hierarchy. If you change the resource hierarchy, the policy hierarchy changes as well. For example, moving a project into an organization will update the project's IAM policy to inherit from the organization's IAM policy.

Child policies cannot restrict access granted at the parent. For example, if you grant editor role to a user for a project, and grant viewer role to the same user for a child resource, then the user still has editor role for the child resource.

## Interacting with Google Cloud Platform

- Three methods of interaction:
  - Cloud Console - Web user interface
  - Cloud SDK/Cloud Shell (command line interface)
  - REST-based API
- Command line options
  - Google Cloud SDK
  - Cloud Shell

## GCP Console

- Easy access to all your Google Cloud Platform projects.
- Access to the Google Cloud Shell.
- A customizable project dashboard, with an overview of Google Cloud resources, billing, and a filterable activity listing.

- Easy access to all Google Cloud Platform APIs, with a dashboard specific to each API, and access to manage your resources.
- Links to Google Cloud Platform starting points, news, and documentation.

## Cloud SDK

- Command line tools for GCP
- Manage resources and applications hosted on Google Cloud Platform.
- Cloud SDK provides the following tools:
  - gcloud
  - bq
  - gsutil

## gcloud

- gcloud is a command-line tool that you can use to perform many common tasks on Google Cloud Platform. You can use gcloud to create and manage:
  - Google Compute Engine virtual machine instances and other resources
  - Google Cloud SQL instances
  - Google Container Engine clusters
  - Google Cloud Dataproc clusters and jobs
  - Google Cloud DNS managed zones and record sets
  - Google Cloud Deployment manager deployments

## bq

bq is a command-line tool that you can use to work with data in Google BigQuery. You can use bq to manage datasets, tables and other entities in BigQuery, as well as run queries on your data.

## gsutil

- gsutil is a command-line tool that you can use to perform tasks in Google Cloud Storage. You can use gsutil to:
- Create and manage Cloud Storage buckets
- Upload objects to buckets, download, and delete them
- Move, copy, and rename objects
- Manage access to stored data

## Cloud Shell

- Google Cloud Shell is an interactive shell environment for Google Cloud Platform. It makes it easy for you to manage your projects and resources without having to install the Google Cloud SDK and other tools on your system. With Cloud Shell, the Cloud SDK gcloud command-line tool and other utilities you need are always available when you need them.
- Cloud Shell provides the following:
  - A temporary Compute Engine virtual machine instance
  - Command-line access to the instance from a web browser
  - 5 GB of persistent disk storage
  - Pre-installed Google Cloud SDK and other tools
  - Language support for Java, Go, Python, Node.js, PHP and Ruby
  - Web preview functionality
  - Built-in authorization for access to Cloud Platform Console projects and resources

## Restful APIs

- Programmatic access to products and services
- Typically use JSON as an interchange format
- Use OAuth 2.0 for authentication and authorization
- Enabled through the Google Cloud Platform Console

- Most APIs include daily quotas and rates (limits) that can be raised by request
- Important to plan ahead to manage your required capacity
- Experiment with APIs Explorer

## APIs Explorer

- The APIs Explorer is an interactive tool that lets you easily try Google APIs using a browser
- With the APIs Explorer, you can:
  - Browse quickly through available APIs and versions.
  - See methods available for each API and what parameters they support along with inline documentation.
  - Execute requests for any method and see responses in real time.
  - Make authenticated and authorized API calls with ease.
  - Support various languages
    - Java, Python, JavaScript, PHP, .NET, Go, Node.js, Ruby, Objective-C, Dart
- https://developers.google.com/apis-explorer

## Cloud Shell Limitations

- 1 hour time out for inactivity
  - Machine will terminate/self-delete
  - $HOME directory contents will be preserved for new session
- Direct interactive use only
  - Not for running high computational/network workloads
  - If in violation, session can be terminated without notice
- For long periods of inactivity, home disk may be recycled (with advance notice via email)
  - If need longer inactive period, consider either local installed SDK or use Google Cloud Storage for long term storage

## Gcloud commands

- For Cloud Shell and locally-installed Google Cloud SDK
- Manage Google Cloud Platform resources and developer workflow
- (Typical) command format:
  - gcloud [GROUP] [GROUP] [COMMAND]—arguments
- Examples:
  - gcloud compute instances create instance-1 –zone us-central1-a
  - gcloud config set project my-unique-project-id
- Full gcloud command list at https://cloud.google.com/sdk/gcloud/reference/

## Cloud Launcher

Google Cloud Launcher lets you quickly deploy functional software packages that run on Google Cloud Platform. Even if you are unfamiliar with services like Compute Engine or Cloud Storage, you can easily start up a familiar software package without having to manually configure the software, virtual machine instances, storage, or network settings.

Deploy a software package now, and scale that deployment later when your applications require additional capacity. Google Cloud Platform updates the images for these software packages to fix critical issues and vulnerabilities but doesn't update software that you have already deployed.

Most packages are free minus normal usage fees.

# Choosing the Correct GCP Compute Service for Your Task

- GCP has several options available for hosting your application on Google Cloud Platform. Each option can take advantage of the entire breadth of services offered by Cloud Platform, including storage, networking, big data products, and Google-grade security.
- Bulk of this course will focus on GCP's Compute options in Compute Engine, App Engine, and Container Engine
- Swiss army knife - choose the right tool for the job
- Your role as a Google Cloud Architect will be to choose the right tool based on the business and technical requirements you have to work with.
- From an application deployment and development standpoint, all of the above options will get the job done and each have their pros/cons, depending on your needs
- There is no right or wrong answer, only which one is correct for your environment
- Options roughly operate on a sliding scale flexibility on one end and managed abstraction on the other end
- Guide: https://cloud.google.com/docs/choosing-a-compute-option
- Blog decision guide: https://cloudplatform.googleblog.com/2017/07/choosing-the-right-compute-option-in-GCP-a-decision-tree.html

# Compute Engine

## What is Infrastructure as a Service?

- One of three cloud computing models, the other two being Platform as a Service (PaaS) and Software as a Service (SaaS)
- It is a form of cloud computing that provides virtualized computing resources over the internet
- Often referred to as the core layer of cloud computing. Behind the scenes, PaaS, and SaaS are running on an IaaS layer.
- Deals with virtual machines and (in some cases) virtual networks
- What is considered to be 'infrastructure' in this case? What goes into a virtual machine?
    - Best definition is a virtual version of a physical PC/server. Just as a physical computer has a CPU, memory, hard drive, an operating system, and firewall to manage network traffic, so does a virtual machine.
    - Like a physical machine, it is up to the customer to manage the above aspects, as well as perform typical OS maintenance such as updates.

## Google Compute Engine Overview

### What is Google Compute Engine?

Google Compute Engine delivers virtual machines running in Google's innovative data centers and worldwide fiber network. Compute Engine's tooling and workflow support enable scaling from single instances to global, load-balanced cloud computing.

## What makes Google Compute Engine unique?

- Industry leading price and performance
    - Compute Engine VMs boot quickly and are consistently high performance.
    - Offers local solid state drive (SSD) performance.
    - Positioned as the higher-performance option
        - “Google Compute Engine ranked #1 in price performance” (https://lp.google-mkto.com/rs/248-TPC-286/images/Cloud-Spectator-Best-Hyperscale-Cloud-Providers.pdf)
- Low Cost, Automatic Discounts
    - Google bills in minute-level increments (with a 10-minute minimum charge), so you only pay for the compute time you use. With sustained use discounts, they automatically give you discounted prices for long-running workloads with no up-front commitment required.
- High speed, secure, private network
    - Global private fiber network
    - Connections between data centers on private high speed connections
    - Same network infrastructure used by Google
- Extremely flexible
    - Create custom images
    - Low cost preemptible VMs for batch workloads
    - Custom machine types (CPU/memory specs)
    - Resize disks with no downtime
    - Create metadata and startup scripts

# Snapshots and Images

- Similarities:
    - Can be used to create a new instance within the same project and in different zones.
- Differences:
    - Intended purpose:
        - Snapshots are different from images, which are used create instances and instance templates. Snapshots are useful for periodic backup of the data on your persistent disks. You can create snapshots from persistent disks even while they are attached to running instances.
        - i.e., backup and disaster recovery
    - Lower cost than custom images
    - Differential backups – only the data changed since the last snapshot is recreated

# Preemptible VM's and groups

## What are Preemptible VM's?

- Short lived, low cost VM
- Costs less than a typical instance (up to 80% cheaper)
- Short lifespan, max of 24 hours
- Excellent for short term batch processing
- Compute-intensive workloads
- Deploy in large numbers as needed
- Fault tolerant workloads

## What are Instance Groups?

- Create and manage groups of virtual machine instances so that you don't have to individually control each instance in your project.
- Managed and unmanaged
- Managed groups:
  - Automatically scale up and down as needed
  - Work with load balancing – distributing traffic among instances
  - If an instance crashes, it is auto-recreated
  - Instance templates define and deploy the group
- Think of it as a 'hive mind' effect. Many machines acting as one.

# Networking

## Virtual Private Cloud (VPC) Network

A VPC network is a virtual version of the traditional physical networks that exist within and between physical data centers. A VPC network provides connectivity for your Compute Engine virtual machine (VM) instances, Container Engine containers, App Engine Flex services, and other network-related resources.

Each GCP project contains one or more VPC networks. Each VPC network is a global entity spanning all GCP regions. This global VPC network allows VM instances and other resources to communicate with each other via internal, private IP addresses.

Each VPC network is subdivided into subnets, and each subnet is contained within a single region. You can have more than one subnet in a region for a given VPC network. Each subnet has a contiguous private RFC1918 IP space. You create instances, containers, and the like in these subnets. When you create an instance, you must create it in a subnet, and the instance draws its internal IP address from that subnet.

Virtual machine instances in a VPC network can communicate with instances in all other subnets of the same VPC network, regardless of region, using their RFC1918 private IP addresses. You can isolate portions of the network, even entire subnets, using firewall rules.

## External IP addressing

A static external IP address is an external IP address that is reserved for your project until you decide to release it. If you have an IP address that your customers or users rely on to access your service, you can reserve that IP address so that only your project can use it. It is also possible to promote an ephemeral external IP address to a static external IP address.

## Routes

A route is a mapping of an IP range to a destination. Routes tell the VPC network where to send packets destined for a particular IP address.

By default, every network has routes that let instances in a network send traffic directly to each other, even across subnets. In addition, every network has a default route that directs packets to destinations that are outside the network. While these routes cover most of your normal routing needs, you can also create special routes that override these routes. For example, you could create a route that forwards packets destined for the Internet to a proxy server first.

## Firewall Rules in GCP

Every GCP network also functions as a managed, distributed firewall. While firewall rules are applied to the network as a whole, the connections are allowed or denied at the instance level. You can think of the firewall as existing not only between your instances and other networks but between individual instances within the same network.

Unless you specify otherwise, GCP applies every rule to every instance in the given network. For example, a firewall rule that denies connections from a range of IP addresses will deny those connections for all instances in the network.

You can restrict which instances a rule applies to by using target tags. For example, you can create a firewall rule that allows only instances marked with a particular tag to reach certain IP ranges.

If all firewall rules in a network are deleted, there is still an implied “Deny all” ingress rule and an implied “Allow all” egress rule for the network.

## Cloud DNS

- DNS translates a computer domain names (like google.com) into IP addresses
- Google Cloud DNS is a high-performance, resilient, global Domain Name System (DNS) service that publishes your domain names to the global DNS in a cost-effective way.
- Create managed zones
    - Add, edit, and delete DNS records

## Cloud VPN

- Google Cloud VPN securely connects your on-premises network to your Google Cloud Platform (GCP) Virtual Private Cloud (VPC) network through an IPsec VPN connection.
- Traffic traveling between the two networks is encrypted by one VPN gateway, then decrypted by the other VPN gateway. This protects your data as it travels over the Internet.
- Supports site to site (gateway to gateway) connections, but not client-gateway connections.

## Cloud Router

- You can use use VPNs to securely connect your on-premises networks your VPC networks.
- Without Cloud Router, you are required to configure your VPNs using static routes. With Cloud Router, your Cloud VPNs support dynamic routing.
- Cloud Router peers with your VPN gateway or router and exchanges topology information via BGP. Network topology changes propagate automatically between your VPC network and your on-premises network, thereby eliminating the need to configure static routes for your Cloud VPN tunnels.
- Cloud Router is a fully distributed and managed Google cloud service that is architected using the principles of Software-Defined Networking (SDN) and delivered on Google's software-defined network virtualization stack.

## Cloud Interconnect

- Connect your infrastructure to Google's network edge with enterprise-grade interconnect
- Connect to Google through carrier partners (7 as of now)
- Equinix, for example, will offer businesses up to a 10 GB connection to Google's cloud services through its Cloud Exchange platform in 15 markets.
    - Why does this matter?
        - Many enterprises that want to move to the cloud want to be able to use these services similar to how they use their on-premise servers.
        - Need a private network with a very fast connection to the cloud provider's data centers and that's what Google now offers them.
- Direct Peering
    - developers can get a blazing fast connection directly to Google's servers in over 70 points of presence in 33 countries. With direct connect, customers can establish their own private links to Google without using an intermediary network.
    - Connect directly to edge network location
    - Works in 70+ locations in 33 countries

- CDN Interconnect
    - Content Delivery Network - Helps your website to serve images faster and help improves load time.

# Tools

## Stackdriver

- Monitoring, logging, and diagnostics
- Google Stackdriver provides powerful monitoring, logging, and diagnostics for cloud operations. By equipping 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.
- Natively monitor GCP, AWS, or a hybrid of both environments
- Combine metrics, logs, and metadata from both platforms into a single viewing environment
- Partner ecosystem and tools to make working with Stackdriver even easier
- Partners include: PagerDuty, BMC, Splunk, and others.
- Stackdriver Monitoring
    - Full-stack monitoring for Google Cloud Platform and Amazon Web Services.
- Stackdriver Logging
    - Real-time log management and analysis.
- Stackdriver Error Reporting
    - Identify and understand your application errors.
- Stackdriver Debugger
    - Investigate your code’s behavior in production.
- Stackdriver Trace
    - Find performance bottlenecks in production.

## Google Cloud Deployment Manager

- Infrastructure Management Service
- Specify all the resources needed for your application in a declarative format using YAML
  - YAML syntax lists each of the resources in your deployment
- Repeatable deployment process
  - By creating configuration files which define the resources, the process of creating those resources can be repeated over and over with consistent results.
- Declarative language
  - Many tools use an imperative approach, requiring the user to define the steps to take to create and configure resources. A declarative approach allows the user to specify what the configuration should be and let the system figure out the steps to take.
  - Greatly simplifies the process
- Focus on the Application
  - The user can focus on the set of resources which comprise the application or service instead of deploying each resource separately.
- Template-Driven
  - Templates allow the use of building blocks to create abstractions or sets of resources that are typically deployed together (e.g. an instance template, instance group and autoscaler). These templates can used over and over by changing input values to define what image to deploy, the zone in which to deploy or how many virtual machines to deploy.

## Google Cloud Source Repositories

- Git repository hosted on GCP
- 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. When paired with Stackdriver Debugger, you can use Cloud Source Repositories and other tools to view debugging information along with code during application runtime. Cloud Source Repositories also provides a source browser to view your repository files from within the Cloud Console.

- Connect to GitHub or Bitbucket
- Built in Source Code Editor
- Integrates with Stackdriver debugger

# Google App Engine

## What is Platform as a Service (PaaS)?

- Definition: PaaS is a category of cloud computing that provides a platform and environment to allow developers to build applications and services over the internet without having to worry about allocating and managing infrastructure.
- Popular with software and web developers
- Key distinctions vs. IaaS:
  - Focused on application development
  - Managed infrastructure
  - Pay per use vs. pay per allocation

## What is App Engine?

- GCP's tool to build modern web and mobile applications on an open cloud platform
- App Engine is a fully-managed application platform that allows developers to build web applications and API services without having to worry about lower-level infrastructure by abstracting away the infrastructure so you focus only on code
- For an 'out of the box' environment with default configurations, App Engine supports Node.js, Java, Ruby, C#, Go, Python, and PHP
- Open Tools that don't lock users in: Developers shy away from proprietary tools that lock them into one platform (aka vendor lock in).
  - App Engine supports custom Docker images, letting developers bring their own custom software stack, making it easy to migrate their application to a different platform.
  - Per Google engineer on vendor lock in: "it's not our strategy – in fact we actively work to minimize it."
- Simply defined:
  - You bring your code, Google handles the rest.

- What does this look like in practice?
  - With traditional physical infrastructure, you need to manage:
    - Firewalls
    - Denial of service attacks
    - Viruses
    - Patches
    - Network configurations,
    - Failover
    - Load balancing
    - Capacity planning (scaling)
    - OS patches and upgrades (in particular security related)
    - Hardware upgrades or fixes
    - Certification levels
    - Most security issues
    - Routing
    - IP addressing
  - IaaS solutions (like Compute Engine) help with several of the above aspects, but App Engine handles all of them for you

# App Engine Environments: Standard vs. Flexible

### Standard

The App Engine standard environment is based on container instances running on Google's infrastructure. Containers are preconfigured with one of several managed available runtimes (Java 7, Java 8, Python 2.7, Go and PHP). Each runtime also includes libraries that support App Engine Standard APIs. For many applications, the standard environment runtimes and libraries might be all you need.

- The App Engine standard environment makes it easy to build and deploy an application that runs reliably even under heavy load and with large amounts of data. It includes the following features:
    - Persistent storage with queries, sorting, and transactions.
    - Automatic scaling and load balancing.
    - Asynchronous task queues for performing work outside the scope of a request.
    - Scheduled tasks for triggering events at specified times or regular intervals.
    - Integration with other Google cloud services and APIs.
- Applications run in a secure, sandboxed environment, allowing App Engine standard environment to distribute requests across multiple servers, and scaling servers to meet traffic demands. Your application runs within its own secure, reliable environment that is independent of the hardware, operating system, or physical location of the server.

### Flexible Environment

Based on Google Compute Engine, the App Engine flexible environment automatically scales your app up and down while balancing the load. Microservices, authorization, SQL and NoSQL databases, traffic splitting, logging, versioning, security scanning, and content delivery networks are all supported natively. In addition, the App Engine flexible environment allows you to customize your runtime and even the operating system of your virtual machine using Dockerfiles.

- Runtimes - The flexible environment includes native support for Java 8/Servlet 3.1/Jetty 9, Python 2.7 and Python 3.5, Node.js, Ruby, PHP, .NET core, and Go. Developers can customize these runtimes or provide their own runtime by supplying a custom Docker image or Dockerfile from the open source community.

- Infrastructure Customization - Because VM instances in the flexible environment are Google Compute Engine virtual machines, you can take advantage of custom libraries, use SSH for debugging, and deploy your own Docker containers.

- Performance - Take advantage of a wide array of CPU and memory configurations. You can specify how much CPU and memory each instance of your application needs and the flexible environment will provision the necessary infrastructure for you.

App Engine manages your virtual machines, ensuring that:

- Instances are health-checked, healed as necessary, and co-located with other services within the project.

- Critical, backwards compatible updates are automatically applied to the underlying operating system.

- VM instances are automatically located by geographical region according to the settings in your project. Google's management services ensure that all of a project's VM instances are co-located for optimal performance.

- VM instances are restarted on a weekly basis. During restarts Google's management services will apply any necessary operating system and security updates.

- You always have root access to Compute Engine VM instances. SSH access to VM instances in the flexible environment is disabled by default. If you choose, you can enable root access to your app's VM instances.

## App Engine locations

App Engine is regional, which means the infrastructure that runs your apps is located in a specific region and is managed by Google to be redundantly available across all the zones within that region.

App Engine is available in the following regions:

- us-central1

- us-east1

- us-east4

- europe-west1

- europe-west2

- asia-northeast1
- australia-southeast1

# Google Cloud Platform Storage Options

## Storage Options Overview

- Google Storage Options = ‘store data here’
- Several storage options available on GCP
- What format is our data? - breakdown
  - Structured vs non-structured
  - Structured = table format - columns and rows = database options
  - Non-structured = Google Cloud Storage
- Database options are broken down further:
  - Relational = SQL
  - Non-Relational = NoSQL
  - Can be VERY roughly defined as NoSQL for scale, SQL for consistency
  - SQL = CloudSQL
  - NoSQL = Datastore, Bigtable
  - Further broken down into whether running analytics, Bigtable for analytics, Datastore for not
- New category in Spanner
  - Previously used internally at Google, not released for everyone else to use
  - Like SQL, it is a relational database
  - Claims advantages of both SQL and NoSQL, called ‘NewSQL’

## Cloud Storage

- Has no knowledge of the structure/order of data that you put in, it just stores it for you as requested regardless of internal structure we set.
    - Files referred to as ‘objects’.
- BigQuery = both a storage and analysis/analytics service

## CloudSQL

- Create instance/region/size
- Database version
    - Select MySQL version
    - Hosted MySQL service, not similar to MySQL, it IS MySQL
- Low maintenance MySQL instance
    - OS management/updates handled for you
- Limited scalability

## NoSQL Options

### Datastore

- Born as an App Engine data repository, but now works outside of it
- Scale and flexible
- Fast and loose (flexible)
- No provisioning resources, true NoOps
- Scales from 0 to TB’s of data

### Bigtable

- Hosted version of Google's own internal Bigtable tech
- HBase was born from BigTable
- Managed service
- For:
    - Storing over a TB of structured data
    - Very high volume of writes
    - Read/write latency in single digit millisecond range with high consistency
    - Easy migration from HBase to a managed cloud service
- Since so low-latency, entire database stored in single zone
- More nodes = more data throughput (and more cost)
- Nodes cost per hour, whether used or not
- Super high scaling
- For big applications/data only. Overkill for smaller applications (minimum nodes) from cost perspective

### Above 3 databases are OPERATIONAL - intended to be part of an application

- BigQuery - analytical database, not operational
- Run SQL queries on TBs of data in seconds

# Google Container Engine

## What are Containers?

- Method of operating system virtualization that lets you run an application plus dependencies in isolated processes called containers
- Simply defined: A container contains the entire runtime environment, including application code, configurations, and dependencies, libraries and other binaries, and packages them into self-contained building blocks
    - Differences in host OSes and underlying infrastructure are abstracted away
- With containers, it is easy to get software to run reliably across different computing environments
    - Developer's laptop to test environment, to staging environment, and to production
    - From physical PC to a VM in private or public cloud
    - Before containers, different software environments can break programs
- Docker is one of the more popular container technologies
- How are containers different from standard VMs (IaaS)?
    - A VM contains an entire operating system packaged with the application
    - A container only runs OS kernel. It is much more lightweight and uses less resources
- Other benefits include:
    - Faster start time since no host OS to boot
    - Smaller size
    - Broken down into easier to manage modules
        - Updating application only required updating the needed module
- Docker - popular container technology, and the type Container Engine uses
    - For this lesson, the term "docker" and container are synonymous

- Imagine containers as containers on a large cargo ship. Each container can be added and removed, stacked, etc. as a self contained entity. Same concept with software containers.

## What is Kubernetes?

- Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications.
- With Kubernetes, you can:
  - Deploy your applications quickly and predictably
  - Scale your applications on the fly
  - Roll out new features seamlessly with no downtime
  - Limit hardware usage to required resources only

## What is Google Container Engine (GKE)?

- It is a managed environment for deploying containerized applications
  - Uses Compute Engine instances and resources
- Falls between Compute Engine and App Engine for managed services vs. granular control
- Google has been running containers in their internal production workloads for over 15 years
- As a managed service, many details are handled for you:
  - Simply set CPU, memory, and storage requirements and Container Engine will provision and manage resources automatically
  - Self-healing - replication strategies, monitoring, and automated repairs result in high reliability and availability
  - Autoscaling - GKE automatically scales up and down based on load to meet demand
  - Runs Kubernetes - an open source container orchestrator that Google invented
  - No vendor lock in - take workloads out of GKE and run anywhere Kubernetes is supported, including your own on-premise services

- Runs a custom OS, called Container-Optimized OS
    - Comes with Docker container runtime and all Kubernetes components needed for easy deployment
    - Automatically updated, no server management necessary
    - Update to newer versions of Kubernetes with no downtime

## GKE Organization

- Container cluster
    - Group of Compute Engine instances running Kubernetes
    - Contains 1 or more node instances and managed Kubernetes master endpoint
    - Central foundation of GKE
        - Nodes, pods, services, and replication controllers all run within cluster
- Kubernetes master
    - Manages the cluster
    - Single endpoint for interacting with your cluster
- Nodes
    - Individual Compute Engine instances
    - Run services to support Docker containers
    - Each node contains one or more pod
- Pods
    - Group of one or more containers
    - Pods share storage and configuration data relating to those containers
    - Pods can contain multiple containers, and multiple pods can exist on each node
    - Pods have a short lifespan and can be deleted and recreated as necessary

- Replication controller
    - Ensures the requested number of pod replicas are always available and running at a given time
    - Automatically adds or removes pods as required to maintain a desired state.
- Services
    - Defines a logical set of pods across nodes and a way to access them using single IP address and port number
    - Services create an abstraction layer that decouples frontend clients from pods that provide backend functions. In this way, clients can work without concerns about which pods are being created and deleted at any given moment.
- Container Registry
    - Not part of the Container cluster but a separate service
    - Secure, private Docker image storage for deployments
    - Push container images to registry for deployment to GKE, GCE, or your own hardware

# Big Data and Machine Learning

## What is Big Data, and Why is It Important?

Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

Big data refers to data that would typically be too expensive to store, manage, and analyze using traditional (relational and/or monolithic) database systems. Usually, such systems are cost-inefficient because of their inflexibility for storing unstructured data (such as images, text, and video), accommodating “high-velocity” (real-time) data, or scaling to support very large (petabyte-scale) data volumes.

For this reason, the past few years has seen the mainstream adoption of new approaches to managing and processing big data, including Apache Hadoop and NoSQL database systems.

## Where Does Big Data Come From?

In the past, most customer data came from structured formats (e.g. bank transactions). Today, the large amount of data that organizations create daily in the form of unstructured online customer interactions completely dwarfs the past structured data formats from just a few years ago.

One big source is the ‘Internet of Things’ (IoT) of interconnected devices and sensors, which has created an exponential increase in the sheer volume of data as text, images, video, and audio.

Also, in some regulated industries, data that in the past would be deemed unimportant and archived now needs to be accessible and analyzed for compliance reasons.

## Why is Big Data Important?

- Business value
    - The ability to get business value out of this mass of data on a consistent basis is now a trait of successful organizations across every industry
    - In some industries such as advertising and retail, it’s literally a matter of survival
- If you can get more data, and if you can properly act on that data, the more value you will receive
- Big data often goes hand in hand with machine learning, as more data can ‘train’ big data models to get even more value.

## Why the Cloud is the Best Platform for Big Data

- Today's big data needs often prove to be complex to deploy, manage, and use in an on-premise situation. The space, costs, and administration requirements necessary to maintain them are high.
- Managed big data services (like BigQuery and Machine Learning Engine) reduce cost and reduce the complexities of hosting your own system
    - Offers the ability to experiment with different products/systems without the up front cost

## How Does GCP Do Big Data Well?

- Big data analytics at Google scale
    - Able to store and process the sheer volume of data
    - Big data is in their DNA
- Serverless, managed infrastructure
    - All backend infrastructure handled for you, including auto scaling
- Fast action on petabyte sized datasets
- Both batch and stream processing
- Spark and Hadoop in the cloud
- Industry leading Machine Learning capabilities

# GCP Big Data Services

- Google BigQuery
    - Google BigQuery is an enterprise data warehouse that stores and queries massive datasets (100's of terabytes) by enabling super-fast SQL queries using the processing power of Google's infrastructure.
    - BigQuery replaces the typical hardware setup for a traditional data warehouse and serves as a collective home for all analytical data in an organization.

        - BigQuery is fully-managed. You don't need to deploy any resources, such as disks and virtual machines.
        - Real time analysis - Real-time analytics is the use of, or the capacity to use, data and related resources as soon as the data enters the system.
- Google Cloud Dataflow
    - Fully managed data processing service
        - No resource provisioning – on demand and nearly limitless capacity
    - Batch and stream processing
        - Batch = process large volume of data at once (e.g. all data from past week)
        - Stream = continuous input and output of data
            - Fast turn-around of data
    - Open source – no vendor lock in
    - Tight integration with rest of GCP
- Google Cloud Dataproc
    - Cloud Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.
    - Scalable clusters - quickly create and resize clusters as needed
- Google Cloud Datalab
    - Cloud Datalab is a powerful interactive tool created to explore, analyze, transform, and visualize data and build machine learning models on Google Cloud Platform. It runs on Google Compute Engine and connects to multiple cloud services easily so you can focus on your data science tasks.
    - Open source – built on Jupyter (formerly iPython)
    - Integrates with other GCP services
    - Supports machine learning models based on TensorFlow

- Google Cloud Dataprep
    - Google Cloud Dataprep is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.
- Google Cloud Pub/Sub
    - Google Cloud Pub/Sub brings the scalability, flexibility, and reliability of enterprise message-oriented middleware to the cloud. By providing many-to-many, asynchronous messaging that decouples senders and receivers, it allows for secure and highly available communication between independently written applications. Google Cloud Pub/Sub delivers low-latency, durable messaging that helps developers quickly integrate systems hosted on the Google Cloud Platform and externally.
    - Send and receive messages between messages in real time
    - Ideal for stream processing
    - Connecting service between other GCP services
    - Decouples senders and receivers
    - Dependable communication between independently written apps
    - For example, a residential sensor can stream data to backend servers hosted in the cloud.

# Machine Learning

## What is Machine Learning?

- New field in computer science in which machines can learn and adapt new things through experience, without having to be explicitly programmed to do so.
    - When exposed to new data, a ML app can learn from it without external input
- Machine learning can be simply defined as creating applications that can **see, hear, and understand** the world around them. This is a new class of intelligent applications.
- Big data and machine learning go hand in hand

## Why Does Machine Learning Matter?

- Machine learning is the next frontier in technology innovation
- All of the major technology companies are making it a focus
- Practical applications include:
    - Predictive protection against cyber-crime, using big data models
    - Automatic identification of images (Google Photos)
    - Real time language translation
        - Text, audio, and even images
    - Voice-to-text dictation and vice versa
    - Smartphone personal assistants - Siri, Google Assistant, Alexa
        - Automated notifications of when to leave for an appointment
    - Automates tasks in analyzing big data
    - In the future - self driving cars

## Google Cloud and Machine Learning

- Google's ML services are faster and more accurate than its competitors
    - Google is by far a leader in ML capabilities
- Uses the same ML resources that Google uses for their own products such as Google Photos, Translate, and Google Assistant
- Built on Tensorflow - Google's Open source tool to build and run neural network models
    - Released in 2015 as an open source ML platform for anyone to use
    - Make it easier for developers to design, build, and train deep learning models
    - Previously used internally at Google for their own deep learning data flows
    - Works on a large variety of platforms

- Create your own machine learning service on GCP's Machine Learning Engine
    - Google Cloud Machine Learning Engine is a managed service that enables you to easily build machine learning models that work on any type of data, of any size.
        - Since it is managed, no resource allocation necessary
    - Ideal for custom predictive analytics
    - Create own model with Tensorflow framework
        - Open source
    - Tightly integrates with Cloud Storage, BigQuery, and other Big Data resources
    - Supports thousands of users and TB's of data
    - Pre-trained machine learning models built by Google
- Vision: understand the content of an image by identifying objects, landmarks, text, explicit content without manually tagging each image
    - Classify image into thousands of categories (boat, lion, car, etc)
    - Detect inappropriate content
    - Find topical web items such as celebrities, logos, and events
    - Detect and extract text for OCR - automatic language detection
- Natural Language: reveal the structure and meaning of text, including language detection
    - Extract information about people, places, events and much more, mentioned in text documents, news articles or blog posts.
    - Understand written sentiment about your product on social media or parse intent from customer conversations happening in a call center or a messaging app.
- Translate: language detection and translation
- Speech: convert audio to text and vice versa
    - Stream results in real-time
    - Over 110 languages supported

- Video Intelligence - newest option, unique to Google
    - Makes videos searchable by content
    - Automatically label objects and people in videos
    - Detect scene transitions and search within a scene

## GCP Machine Learning Use Cases

- Structured data
    - Forecasting customer demand
    - Product upsells on product website
    - Fraud detection - detect anomalies in data
- Unstructured data
    - Language identification
    - With image detection - Inspect packages for damage