GCP ACE:

INTRODUCTION:

Module Overview-

[00:00](javascript:;)>> Welcome to preparing for your associate cloud engineer journey.

[00:04](javascript:;)In this course, you'll learn more about the skills covered on the associate cloud engineer certification exam.

[00:10](javascript:;)It's important to clarify that this course by itself will not prepare you to take the A certification exam.

[00:17](javascript:;)This is not a cram session.

[00:19](javascript:;)The exam is purposely calibrated to test your ability to apply the knowledge required of an associate cloud engineer, not merely repeated.

[00:28](javascript:;)Cram sessions have minimal impact on your ability to pass the exam.

[00:32](javascript:;)Instead, the goal of this course is to help you structure your preparation time for the exam.

[00:38](javascript:;)You'll learn about the scope of each exam section, assess your current knowledge and skills through

[00:43](javascript:;)diagnostic questions, and review where to find additional tools and resources to include in your study plan.

[00:50](javascript:;)In this introductory module, you'll learn about the role of an associate cloud engineer, the types of resources available

[00:56](javascript:;)to support your study, and how you will use the workbook in this course to create your study plan.

Introduction to the Associate Cloud Engineer Role-

[00:00](javascript:;)Person: So you're preparing for the Associate Cloud Engineer certification.

[00:04](javascript:;)But what exactly is an Associate Cloud Engineer?

[00:07](javascript:;)Let's review the job-role description.

[00:08](javascript:;)An Associate Cloud Engineer deploys and secures applications and infrastructure, monitors operations of multiple projects and maintains enterprise solutions to ensure that they meet target performance metrics.

[00:23](javascript:;)This individual has experience working with public clouds and on-premises solutions.

[00:28](javascript:;)They are able to use Google Cloud Console and the command line interface to perform common platform based

[00:34](javascript:;)tasks to maintain and scale one or more deployed solutions that leverage Google-managed or self-managed services on Google Cloud.

[00:42](javascript:;)In this course, you'll examine the role of an Associate Cloud Engineer by putting yourself in the shoes of an Associate

[00:48](javascript:;)Cloud Engineer at Cymbal Superstore, a fictional company that is preparing to migrate some of its on-premises applications to Google Cloud.

[00:57](javascript:;)Cymbal Superstore is experiencing several challenges common to legacy retailers.

[01:03](javascript:;)Customers increasingly expect a seamless online offline experience when they shop, making the Cymbal Superstore e-commerce application vital.

[01:13](javascript:;)Cymbal Superstore logistics wants to use data to more efficiently oversee transportation.

[01:18](javascript:;)Cymbal Superstore operations has been grappling with how to update their supply chain strategy to meet customer demand in a timely, efficient way.

[01:28](javascript:;)To meet these challenges, Cymbal Superstore plans to migrate their existing e-commerce, supply chain and transportation management applications to Google Cloud.

[01:38](javascript:;)As an Associate Cloud Engineer at Cymbal Superstore, your role involves helping set up the cloud environment.

[01:45](javascript:;)You're also involved in planning, configuring and managing the cloud solutions.

[01:51](javascript:;)This requires you to be familiar with variety of compute, storage and networking resources in Google Cloud.

[01:58](javascript:;)You also need a solid understanding of Google recommended practices for configuring access to cloud resources.

[02:06](javascript:;)As you continue through this course, you'll explore the role of an Associate Cloud Engineer as Cymbal Superstore migrates their applications to Google Cloud.

[02:16](javascript:;)We'll use this scenario to illustrate the types of configurations and tasks that correspond to each section of the exam guide.

[02:24](javascript:;)Cymbal Superstore's Cloud solutions will also provide context for many of the diagnostic questions you'll encounter along the way.

Certification Value and Benefits-

[00:00](javascript:;)>> Why become a Google Cloud Certified Associate Cloud Engineer?

[00:04](javascript:;)Becoming Google certified gives you industry recognition.

[00:07](javascript:;)It validates your technical expertise and can be the starting point to take your career to the next level.

[00:14](javascript:;)Certification value has skyrocketed.

[00:17](javascript:;)The associate cloud engineer certification is valuable on its own, and can also be used as a starting point on the path to professional certification.

[00:25](javascript:;)You may be curious about what separates an associate cloud certification from a professional cloud certification.

[00:33](javascript:;)There is some overlap between the two levels of certification.

[00:36](javascript:;)However, only the professional level certification expects the exam taker to know how to evaluate case studies and design solutions to meet business requirements.

[00:47](javascript:;)The associate level is mainly concerned with technical requirements and customer implementation.

[00:53](javascript:;)That is where you should concentrate your efforts.

Certification Process-

[00:00](javascript:;)Person: The Associate Cloud Engineer Certification Exam is based on the exam guide.

[00:04](javascript:;)In the following modules, you'll take diagnostic questions to assess your knowledge of each section of the exam guide.

[00:11](javascript:;)The exam guide is divided into five sections.

[00:14](javascript:;)Each section has several objectives.

[00:17](javascript:;)We'll focus on where you can find resources at the section objective level.

[00:22](javascript:;)You can find the exam guide on the certification page at the URL shown on the screen.

[00:26](javascript:;)Throughout this course, you'll be pointed to specific resources and documentation that can help you fill the gaps you identify through the diagnostic questions.

[00:36](javascript:;)Let's go over the types of resources you may want to include in your study plan.

[00:41](javascript:;)Google provides resources to help you develop your skills and experience with Google Cloud Solutions.

[00:47](javascript:;)The learning path for this certification includes in-person or online courses, online practice labs and quests and practice questions.

[00:56](javascript:;)The courses recommended for the Associate Cloud Engineer certification include Google Cloud Fundamentals Core Infrastructure, Architecting with Google Compute Engine and Getting Started with Google Kubernetes Engine.

[01:09](javascript:;)You'll learn more about how these courses relate to the sections of the exam guide as you complete the modules in this course.

[01:16](javascript:;)It's important to note that you can take these as instructor-led courses or as on-demand courses on Coursera, Pluralsight or Google's Cloud Skills Boost.

[01:25](javascript:;)The on-demand content included in Architecting with Google Compute Engine is presented over three courses, Essential Google

[01:33](javascript:;)Cloud Infrastructure Foundation, Essential Google Cloud Infrastructure Core Services and Elastic Google Cloud Infrastructure Scaling and Automation.

[01:43](javascript:;)The skills badges provide hands-on experience working in Google Cloud.

[01:48](javascript:;)They consist of a series of related labs followed by a challenge lab to complete the quest and are available on Cloud Skills Boost.

[01:56](javascript:;)As we review the diagnostic questions in this course, you'll also get recommendations for skill badges to include in your study plan.

[02:03](javascript:;)Sample questions are another resource you can use to prepare.

[02:07](javascript:;)The diagnostic questions in this course are designed to help you identify your knowledge gaps.

[02:12](javascript:;)On the certification page, Google provides a different set of sample questions that can help you familiarize yourself with the format of the exam questions.

[02:21](javascript:;)Once you complete the question set, you will receive feedback describing the rationale for the correct answers.

[02:29](javascript:;)The sample questions provide a good opportunity to practice taking the type of scenario-based application-level questions on the exam.

[02:37](javascript:;)The exam questions present you with a scenario, explain the goal or what you are trying to achieve and ask you what you would do in this situation.

[02:46](javascript:;)Remember these tips for multiple-choice questions.

[02:50](javascript:;)Read the question stem carefully.

[02:53](javascript:;)Make sure you understand exactly what the question is asking.

[02:57](javascript:;)Try to anticipate the correct answer before looking at the options.

[03:01](javascript:;)You should be able to come up with the correct answer just from read the question stem.

[03:05](javascript:;)You may find that more than one answer may be possible on multiple-choice tests.

[03:12](javascript:;)Take questions at face value.

[03:13](javascript:;)If certain details are omitted, then they are unlikely to contribute to the selection of the best answer.

[03:21](javascript:;)Pay attention to qualifiers such as usually, all, never, none and key words like the best, the least and except.

[03:32](javascript:;)Google also supplies official public documentation for its products and services.

[03:38](javascript:;)This documentation can be find at the URL shown on the screen.

[03:42](javascript:;)In each of the following modules, you'll learn about specific documentation resources to help you study that section in preparation for the exam.

Creating Your Study Plan-

[00:00](javascript:;)>> One of the primary goals for this course is to help you devise a study strategy that focuses on areas you need to work on.

[00:07](javascript:;)Let's quickly explore how the course is set up.

[00:09](javascript:;)The course and your course workbook focuses on each section of the Exam Guide in turn.

[00:15](javascript:;)To help you craft a study strategy, you'll take diagnostic questions as part of each module.

[00:21](javascript:;)Many of these questions relate to our simple superstore scenario, and ask you to apply concepts you will need to be familiar with as an Associate Cloud Engineer.

[00:31](javascript:;)You can take these questions either using Google Forms or using your workbook.

[00:36](javascript:;)Keep in mind that these diagnostic questions are meant to help you identify gaps in your knowledge, but they don't represent all possible topics on the exam.

[00:46](javascript:;)Remember, we don't expect that you'll answer all these questions correctly right now.

[00:52](javascript:;)This is meant to be a course that you take towards the beginning of your Associate Cloud Engineer journey.

[00:58](javascript:;)And many of you may not be Google Cloud experts yet.

[01:01](javascript:;)We'll review the answers to the questions related to each section objective.

[01:06](javascript:;)As we cover each objective, you'll learn more about where the key concepts appear in the Google Cloud documentation, specific courses and modules, and/or specific skill badges.

[01:17](javascript:;)At the end of each section objective, you'll find a list of related resources, mark or highlight the specific resources you need to add to your study plan.

[01:27](javascript:;)In the final part of your workbook, you'll find a template to help you identify weekly goals and study activities.

[01:34](javascript:;)We'll talk more about putting together weekly goals at the end of this course.

[01:39](javascript:;)Now that you know about the overall setup of this course and how to use the workbook, let's get started by exploring Section 1 of the Exam Guide.

SETTING UP A CLOUD SOLUTION ENVIRONMENT:

Module Overview-

[00:02](javascript:;)>> Welcome to Module 1: Setting Up a Cloud Solution Environment.

[00:06](javascript:;)In this module, you'll explore the scope of task involved in setting up a cloud solution environment, which corresponds to the first section of the Associate Cloud Engineer exam guide.

[00:18](javascript:;)We'll start by discussing some considerations for how you would set up Cymbal Superstore's cloud environment.

[00:25](javascript:;)Next, you'll assess your skills in this section through 10 diagnostic questions, then we'll review these questions.

[00:33](javascript:;)Based on the areas you need to learn more about, you'll identify resources to include in your study plan.

Setting Up Cymbal Superstore’s Cloud Environment-

[00:00](javascript:;)Person: Let's start with how to set up Cymbal Superstore's cloud environment.

[00:04](javascript:;)Cymbal Superstore is ready to start implementing the basic cloud infrastructure for their organization.

[00:11](javascript:;)As an associate cloud engineer, you have been assigned to the team to help with this phase of the project.

[00:17](javascript:;)The team begins with a planning meeting to decide how Cymbal Superstore's cloud structure will be organized.

[00:25](javascript:;)Outcomes of this phase of the migration journey include establishing a resource hierarchy, implementing organizational policies, managing projects and quotas, managing users and groups, and applying access management.

[00:42](javascript:;)Setting up billing and monitoring the use of your cloud resources are also things to consider.

[00:48](javascript:;)Finally, choosing how you interact with Google Cloud is an important decision as well.

[00:53](javascript:;)Different options include the Graphical Cloud Console, command line tools, and software development tools.

[01:00](javascript:;)Let's take a closer look.

[01:03](javascript:;)How to set up a resource hierarchy on Google Cloud depends on the needs and structure of the organization.

[01:09](javascript:;)Cymbal Superstore plans to migrate three applications to the cloud; its supply chain application, which will run on virtual machines, its

[01:18](javascript:;)containerized Ecommerce application, which will use GKE, and its transportation management application, which will use cloud functions to track truck location.

[01:32](javascript:;)These applications correspond to three different departments; Operations Managers, the B2B supply chain, logistic overseas transportation, and the Sales and Marketing Department manages the Ecommerce application.

[01:46](javascript:;)The graphic shows a possible organization hierarchy for Cymbal Superstore.

[01:52](javascript:;)The Cymbal Superstore organization is at the top, followed by optional folders underneath, one for each division and one for each application.

[02:01](javascript:;)There are multiple projects under the B2B supply chain app folder, one for each environment related to Continuous Integration and Continuous Delivery, CICD, a developing, staging and production environment.

[02:15](javascript:;)Other apps in the chart would have a similar project structure.

[02:20](javascript:;)As you set up the hierarchy, you will also need to grant organizational policies.

[02:25](javascript:;)For example, Marketing might need access to data in the supply chain system to see if inventory levels affected recent marketing campaigns.

[02:35](javascript:;)Giving Marketing the viewer permissions on the supply chain production environment might be a good choice to give them the access they need.

[02:44](javascript:;)An associate cloud engineer could also be tasked with enabling Application Programming Interfaces, or APIs, within projects during setup.

[02:53](javascript:;)Cymbal Superstore's Ecommerce project requires access to Google Kubernetes and Cloud SQL as a database backend, and you need to enable the APIs for these services.

[03:07](javascript:;)Setting up the cloud environment also involves granting members IM roles to ensure they have the

[03:13](javascript:;)right access to projects, depending on the needs of their job and their role at Cymbal Superstore.

[03:20](javascript:;)For example, data analysts in the Marketing Department will need access to historic sales data in the Ecommerce system.

[03:28](javascript:;)This is a great use case for BigQuery, Google's cloud-based data warehouse solution.

[03:34](javascript:;)How would we go about giving the proper access?

[03:38](javascript:;)New data analysts would be added to a group named "Data Analyst."

[03:43](javascript:;)Managing permission and roles at a group level is easier than keeping track of permissions for individual users.

[03:51](javascript:;)Data analysts need access to the data in a data center table, so they would require BigQuery.dataviewer role at the proper level.

[04:01](javascript:;)Queries in BigQuery are executive as an executable job.

[04:06](javascript:;)So to submit a query, a data analyst would also need the BigQuery.jobs.create permission.

[04:15](javascript:;)This permission is included in the pre-defined role named BigQuery.user.

[04:22](javascript:;)You would give the data analyst group access to this role in the production project.

[04:29](javascript:;)As a cloud engineer, you'll need to know how to manage Cymbal Superstore's users and groups in

[04:34](javascript:;)cloud identity, a service for managing users and groups, if you're not doing so, via Google Workspace.

[04:43](javascript:;)Google Cloud's operations suite, which used to be called stack driver, provides metrics and logging services for all your services, resources and projects in your cloud environment.

[04:56](javascript:;)To monitor metrics for multiple projects, you set up project scoping.

[05:02](javascript:;)If Cymbal Superstore's operations department decides to monitor metrics across all three supply chain projects in the staging

[05:10](javascript:;)environment project, you will set staging as a scoping project and then add Dev and Production as monitored projects.

[05:22](javascript:;)While migrating to Google Cloud, Cymbal Superstore will be moving some of its IT expenditure to operational

[05:29](javascript:;)expenses, and the different departments associated with each application will be responsible for compute and storage costs.

[05:38](javascript:;)You'll need to create a different billing account for each group and link each project to the appropriate account.

[05:46](javascript:;)The department lead in Sales and Marketing expresses particular concern about the cost of housing its data warehouse in BigQuery, and how to optimize queries and storage.

[05:58](javascript:;)You'll need to set up custom billing projects and alerts for this department.

[06:04](javascript:;)Each department will also need you to set up billing exports that can be used to track charges.

[06:11](javascript:;)As an IT administrator for Cymbal Superstore, you can always set up resources in the graphical console.

[06:18](javascript:;)If you have a combination of tasks you need to do often, you can automate resource management tasks on the command line.

[06:25](javascript:;)The Cloud SDK has a G-Cloud command set that allows configuration of Google Cloud resources as an executable script.

[06:34](javascript:;)You can use the command "gcloud config set" to configure default options, such as the project and compute region.

[06:43](javascript:;)Then if you don't give specific command line arguments, the SDK will use these defaults, which simplifies your code even more.

Introduction: Diagnostic Questions-

[00:00](javascript:;)Person: Now it's your turn to assess your experience and skills in this section with some diagnostic questions.

[00:06](javascript:;)Remember these questions are intended to help you understand or diagnose which areas you'll want to

[00:11](javascript:;)focus on in your study plan, so we don't expect you to know all the answers yet.

[00:18](javascript:;)Please take 15 minutes to complete the diagnostic questions for this section.

Diagnostic Questions-  
The Operations Department at Cymbal Superstore wants to provide managers access to information about VM usage without allowing them to make changes that would affect the state. You assign them the Compute Engine Viewer role. Which two permissions will they receive?

checkcompute.images.list  
compute.images.get

What Google Cloud project attributes can be changed?

The Project Name.

How are resource hierarchies organized in Google Cloud?

Organization, Folder, Project, Resource

Correct! Organization sits at the top of the Google Cloud resource hierarchy. This can be divided into folders, which are optional. Next, there are projects you define. Finally, resources are created under projects.

How are billing accounts applied to projects in Google Cloud? (Pick two).

A billing account can be linked to one or more projects.

A project and its resources can only be tied to one billing account.

You need to add new groups of employees in Cymbal Superstore’s production environment. You need to consider Google’s recommendation of using least privilege. What should you do?

Grant predefined and custom roles that provide necessary permissions and grant basic roles only where needed.

Correct! Basic roles are broad and don’t use the concept of least privilege. You should grant only the roles that someone needs through predefined and custom roles.

You want to use the Cloud Shell to copy files to your Cloud Storage bucket. Which Cloud SDK command should you use?

Gsutil

Pick two choices, from the options below, that provide a command line interface to Google Cloud.

Cloud SDK

Cloud Shell

Fiona is the billing administrator for the project associated with Cymbal Superstore’s eCommerce application. Jeffrey, the marketing department lead, wants to receive emails related to budget alerts. Jeffrey should have access to no additional billing information. What should you do?

Use Cloud Monitoring notification channels to send Jeffrey an email alert.

Correct! You can set up to 5 Cloud Monitoring channels to define email recipients that will receive budget alerts.

Stella is a new member of a team in your company who has been put in charge of monitoring VM instances in the organization. Stella will need the required permissions to perform this role. How should you grant her those permissions?  
Add Stella to a Google Group in your organization. Bind that group to roles/compute.viewer.

Correct! Best practice is to manage role assignment by groups, not by individual users.

Jane will manage objects in Cloud Storage for the Cymbal Superstore. She needs to have access to the proper permissions for every project across the organization. What should you do?

Add Jane to a group that has the roles/storage.objectAdmin role assigned at the organizational level.

Correct! This would give Jane the right level of access across all projects in your company.

Your Study Plan-

[00:00](javascript:;)>> Now let's review how to use these diagnostic questions, to help you identify what to include in your study plan.

[00:07](javascript:;)As a reminder, this isn't meant to be a crash course teaching you everything you need to know about setting up a solution in Google Cloud.

[00:15](javascript:;)Instead, it's meant to give you a better sense of the scope of the section, and the different skills you'll want to develop as you prepare for the certification.

[00:25](javascript:;)We'll approach this review by looking at the objectives of this exam section, and the questions you just answered about this section.

[00:33](javascript:;)We'll introduce an objective, briefly review the answers to the related questions, and then talk about

[00:39](javascript:;)where you can find out more in the learning resources or in the Google Cloud documentation.

[00:46](javascript:;)As we go through each section objective, use the page in your workbook to mark the

[00:51](javascript:;)specific documentation, courses or modules, and quests that you'll want to emphasize in your study plan.

Knowledge Check-

What is the lowest level basic role that gives you permissions to change resource state?

Correct! Editor gives you permissions to change state.

Which Google Cloud interface allows for scripting actions in a set of command line executables?

Correct! Cloud Shell provides a containerized command line interface that allows you to run commands and automate them with scripts.

PLANNING AND CONFIGURING A CLOUD SOLUTION:

Module Overview-

[00:02](javascript:;)>> Welcome to Module 2: Planning and Configuring Cloud Solutions.

[00:04](javascript:;)In this module, you'll explore the scope of tasks involved in planning and configuring Cymbal

[00:09](javascript:;)Superstore's cloud solutions, which corresponds to the second section of the Associate Cloud Engineer exam guide.

[00:16](javascript:;)We'll start by discussing some considerations for how you would decide the kinds of resources needed for Cymbal Superstore's application requirements.

[00:25](javascript:;)Next, you'll assess your skills in this section through 10 diagnostic questions, then we'll review these questions.

[00:32](javascript:;)Based on the areas you need to learn more about, you'll identify resources to include in your study plan.

Selecting Resources for Cymbal Superstore’s cloud solutions-

[00:00](javascript:;)Person: Let's discuss how to evaluate Cymbal Superstore's existing applications so we can plan and configure a cloud solution for them.

[00:07](javascript:;)The first planning meeting for Cymbal Superstore's migration to Google Cloud was a success.

[00:13](javascript:;)Everyone is excited to get into the details of our Cymbal Superstore's existing applications are going to be migrated to the cloud.

[00:20](javascript:;)As an associate cloud engineer, you are tasked to help realize the cloud architect's design during the next phase of Cymbal Superstore's cloud migration, planning and configuring the solution.

[00:32](javascript:;)Compute options in Google Cloud include those based on virtual machines as well as those based on containers.

[00:38](javascript:;)There are also resource-based and several options available from a compute perspective, depending on your team's expertise and focus.

[00:47](javascript:;)The level of controller flexibility needed by your solution could also be factors in deciding the correct compute product to use in a solution.

[00:55](javascript:;)Similarly, data solutions in Google Cloud are based on the needs of the application.

[01:00](javascript:;)Throughput and latency will determine how quickly your application responds.

[01:05](javascript:;)There are some common questions to ask while you are analyzing your data needs.

[01:10](javascript:;)Are your data processing needs transactional or analytical?

[01:14](javascript:;)Do you need to query your data in a relational way?

[01:18](javascript:;)Do you want large groups of related data to be returned through a non-relational Get operation?

[01:24](javascript:;)Your applications require network connectivity is another important facet of cloud solution planning.

[01:31](javascript:;)Does it require access to the internet, or do you just need to provide connectivity to internal components on your private network?

[01:39](javascript:;)Are you going to configure the connectivity of multiple application servers by load balancing incoming traffic across them?

[01:47](javascript:;)How are you going to protect your applicants' network or system outages?

[01:51](javascript:;)Let's take a closer look at how these considerations affect Cymbal's choices.

[01:56](javascript:;)Cymbal Superstore has three applications to migrate to Google Cloud; first, Cymbal Superstore's Sales and Marketing

[02:03](javascript:;)has an existing web application that provides an interface for customers to look at and order products.

[02:10](javascript:;)This Ecommerce app is currently based on containers and needs to have global availability and low latency.

[02:18](javascript:;)Second, delivery services is becoming an important aspect of Cymbal Superstore's customer interactions.

[02:25](javascript:;)CS would like to use Google Services to keep track of truck location.

[02:30](javascript:;)Logistics has a transportation management app implemented locally through a custom message broker.

[02:36](javascript:;)They would like to use Google Cloud to monitor the location of trucks for delivery status and analysis of mileage for preventative maintenance.

[02:45](javascript:;)Third, Operations has decided to migrate their legacy supply chain application to the cloud.

[02:52](javascript:;)This is currently a virtual machine-based application based in a LINUX operating system, and is implemented with a lamp stack.

[03:00](javascript:;)The application needs to be available local to CS queue.

[03:06](javascript:;)Cymbal Superstore's Ecommerce application is currently implemented with containers, so you decide to us Google Kubernetes Engine, of GKE, for compute.

[03:16](javascript:;)For storage, you select Cloud Spanner to house Cymbal Superstore's Ecommerce data, because it allows for global availability and low latency.

[03:26](javascript:;)As an associate cloud engineer, you assist Cymbal Superstore's cloud architect in making these decisions and determining the appropriate configuration.

[03:36](javascript:;)Seeing as the Ecommerce system is a web-based application, you will need an external HTTPS load balancer.

[03:44](javascript:;)With GKE, you can do this by implementing an ingress object with a GCE ingress class.

[03:51](javascript:;)Applying that manifest will create an external load balancer for you.

[03:56](javascript:;)As an associate cloud engineer, you should be familiar with this type of load balancing in Google Cloud.

[04:02](javascript:;)Let's move on to your role in planning and configuring Cymbal Superstore's next application.

[04:07](javascript:;)Logistics transportation management system currently uses an on-premise message broker.

[04:12](javascript:;)Together with the cloud architect, you choose Pub/Sub as a solution to collect sensor data from trucks for analysis in the cloud.

[04:20](javascript:;)You will use cloud functions to pull data from Pub/Sub and start a dataflow pipeline.

[04:25](javascript:;)Data flow will feed data into Cloud Bigtable to store your sensor data.

[04:28](javascript:;)You can run [indistinct] queries in BigQuery to visualize your data in Looker.

[04:34](javascript:;)As an associate cloud engineer, you'll need help to configure all these resources.

[04:40](javascript:;)Moderating and analysis of this data is done close to headquarters.

[04:44](javascript:;)So Cymbal opts for a regional VPC solution for this application.

[04:49](javascript:;)Finally, you play a role in helping configure Cymbal Superstore's third application.

[04:54](javascript:;)Cymbal Superstore's existing supply chain application is implemented using virtual machines with a LINUX operating system and a common LINUX, Apache, MySQL, PHP -- in other words, LAMP development stack.

[05:08](javascript:;)After analyzing this application's need, the cloud architect determines that the recommended cloud solution is to migrate this application to compute engine virtual instances.

[05:19](javascript:;)The data solution uses a cloud SQL instance configured with a MySQL database.

[05:25](javascript:;)The recommended network access would be external HTTPS access to the region for partners and internal connectivity between the application and the backing database service.

Introduction: Diagnostic Questions-

[00:00](javascript:;)>> Now it's your turn to assess your experience and skills related to this section with some diagnostic questions.

[00:06](javascript:;)Remember, the purpose of these questions is to help you better understand what is involved in this

[00:10](javascript:;)section of the Exam Guide and identify which areas you'll want to focus on in your study plan.

Diagnostic Questions:

Which Google Cloud load balancing option runs at Layer 7 of the TCP stack?

Global http(s)

Cymbal Superstore needs to analyze whether they met quarterly sales projections. Analysts assigned to run this query are familiar with SQL. What data solution should they implement?

Correct! BigQuery is Google Cloud’s implementation of a modern data warehouse. BigQuery analyzes historical data and uses a SQL query engine.

The projected amount of cloud storage required for Cymbal Superstore to enable users to post pictures for project reviews is 10 TB of immediate access storage in the US and 30 TB of storage for historical posts in a bucket located near Cymbal Superstore’s headquarters. The contents of this bucket will need to be accessed once every 30 days. You want to estimate the cost of these storage resources to ensure this is economically feasible. What should you do?

Use the pricing calculator to estimate the price for 10 TB of multi-region Standard storage, 30 TB for regional Nearline, and egress charges for reads from the bucket.

Cymbal Superstore decides to migrate their supply chain application to Google Cloud. You need to configure specific operating system dependencies. What should you do?

Implement an application using virtual machines on Compute Engine.

Correct! Compute Engine gives you full control over operating system choice and configuration.

Cymbal Superstore has a need to populate visual dashboards with historical time-based data. This is an analytical use-case. Which two storage solutions could they use? <https://cloud.google.com/bigtable/docs/overview#other-storage-options>

Correct! BigQuery is a data warehouse offering optimized to query historical time-based data. BigQuery can run queries against data in its own column-based store or run federated queries against data from other data services and file stores.

Correct! Cloud Bigtable is a petabyte scale, NoSQL, column family database with row keys optimized for specific queries. It is used to store historic, time-based data and answers the need for this requirement.

You want to deploy a microservices application. You need full control of how you manage containers, reliability, and autoscaling, but don’t want or need to manage the control plane. Which compute option should you use?

Correct! Google Kubernetes Engine gives you full control of container orchestration and availability.

An application running on a highly-customized version of Ubuntu needs to be migrated to Google Cloud. You need to do this in the least amount of time with minimal code changes. How should you proceed?

Create Compute Engine Virtual Machines and migrate the app to that infrastructure

Correct! Compute Engine is a great option for quick migration of traditional apps. You can implement a solution in the cloud without changing your existing code.

Cymbal Superstore decides to pilot a cloud application for their point of sale system in their flagship store. You want to focus on code and develop your solution quickly, and you want your code to be portable. How do you proceed?

Package your code to a container image and post it to Cloud Run.

Correct! Cloud Run provides serverless container management. It lets you focus on code and you can deploy your solution quickly.

Cymbal Superstore’s supply chain application frequently analyzes large amounts of data to inform business processes and operational dashboards. What storage class would make sense for this use case?

Correct. Standard storage is best for data that is frequently accessed ("hot" data) and/or stored for only brief periods of time. In addition, co-locating your resources by selecting the regional option maximizes the performance for data-intensive computations and can reduce network charges.

Cymbal Superstore is piloting an update to its ecommerce app for the flagship store in Minneapolis, Minnesota. The app is implemented as a three-tier web service with traffic originating from the local area and resources dedicated for it in us-central1. You need to configure a secure, low-cost network load-balancing architecture for it. How do you proceed?  
Configure a standard tier proxied external https load balancer connected to the web tier as a frontend and a regional internal load balancer between the web tier and the backend.

Your Study Plan-

[00:00](javascript:;)>> Now let's review the diagnostic questions, paying attention to the areas where you need to focus your study time.

[00:07](javascript:;)Some of this may be content you're already very familiar with, while other aspects may be new to you.

[00:14](javascript:;)Just like we did with the previous module, we'll approach this review by looking at the objectives of this exam section, and the questions you just answered about each one.

[00:24](javascript:;)We'll introduce an objective, briefly review the answers to the related questions, and then talk

[00:30](javascript:;)about where you can find out more in the learning resources and-or in Google Cloud documentation.

[00:37](javascript:;)As we go through each section objective, use the page in your workbook to mark

[00:42](javascript:;)the specific documentation, courses and modules, and quests you'll want to emphasize in your study plan.

Knowledge Check-

Which storage class is designed for long term storage has a 365 day minimum storage agreement, and a lower storage price as compared to other storage types?

Archive storage

Which serverless option is based on developing and executing small snippets of code?

Correct! Cloud Functions allows you to submit and execute small snippets of code that fire based on system events.

DEPLOYING AND IMPLEMENTING A CLOUD SOLUTION:

Module Overview-

[00:01](javascript:;)Person: Welcome to module three, Deploying and Implementing Cloud Solutions.

[00:06](javascript:;)In this module, you'll explore the scope of task involved in deploying and implementing Cymbal Superstore's recommended cloud solutions.

[00:15](javascript:;)These tasks correspond to the third section of the Associate Cloud Engineer exam guide.

[00:21](javascript:;)We'll start by discussing how to deploy and implement Cymbal Superstore's cloud solutions.

[00:27](javascript:;)Next, you'll assess your skills in this section through 10 diagnostic questions.

[00:31](javascript:;)Then we'll review these questions.

[00:35](javascript:;)Based on the areas you need to learn more about, you'll identify resources to include in your study plan.

Deploying and Implementing Cymbal Superstore’s cloud recommended solutions-

[00:00](javascript:;)Person: Way to go!

[00:01](javascript:;)You've planned and configured cloud solutions for Cymbal Superstore's application requirements.

[00:07](javascript:;)It's time to think about how you can deploy and implement the resources needed to realize the company's goals.

[00:15](javascript:;)You've worked hard to make sure resource entities and policies are set up correctly for Cymbal Superstore's cloud architecture.

[00:23](javascript:;)You've also selected cloud products for the application Cymbal Superstore has decided to migrate to the cloud.

[00:30](javascript:;)Solution deployment is a critical part of your role.

[00:35](javascript:;)As an associate cloud engineer, you're expected to have the knowledge to implement specific compute solutions, including Compute Engine, Kubernetes Engine, Cloud Run and Cloud Functions.

[00:49](javascript:;)Understanding availability, concurrency, connectivity and access options for these services are keys to success as you deploy them to support your needs.

[01:01](javascript:;)Solutions you implement in Google Cloud will also require data stores.

[01:05](javascript:;)Google Cloud's data solutions include products that utilize relational and no SQL data structures.

[01:13](javascript:;)There are different products that support transactional and analytical use cases.

[01:18](javascript:;)Some solutions are optimized for low latency and global availability.

[01:23](javascript:;)Properly implementing software-defined networking will ensure your application front ends are accessible, and your application back ends are secured.

[01:34](javascript:;)Cloud Marketplace is always a good place to go if there's a specific software stack you need to support on your Compute Engine instances.

[01:43](javascript:;)With Cloud Marketplace, you don't have to reinvent the wheel and can have a development or production framework up and running in no time.

[01:52](javascript:;)A common dev op's practice is to deploy your infrastructure in a declarative way, and source control your configuration files.

[02:02](javascript:;)Deploying resources via infrastructure as code reduces human error, and speeds up resource allocation.

[02:09](javascript:;)Knowing how to do this in the context of your role as an associate cloud engineer is yet another tool you have at your disposal.

[02:18](javascript:;)As I review, here are Cymbal Superstore's proposed solutions.

[02:24](javascript:;)Their Ecommerce solution is based on container management provided by Google Kubernetes Engine.

[02:28](javascript:;)Data provided by the globally available, horizontally scalable capabilities of Cloud Spanner, and external HTTPS load balancing for user access.

[02:41](javascript:;)This use case also has a need for historical sales data to be analyzed by BigQuery, Google Cloud's modern data warehouse implementation.

[02:52](javascript:;)The transportation management cloud solution monitors Pub/Sub for incoming sensor data, triggers a cloud function as new messages

[03:01](javascript:;)are posted to a specific topic, and starts a dataflow job to transform data and save it into Bigtable.

[03:10](javascript:;)Finally, the supply chain application implements managed instance groups in Compute Engine.

[03:17](javascript:;)The back end store for this solution is Cloud SQL.

[03:21](javascript:;)Connectivity between the back end database and Compute Engine instances is via TCP internal to the VPC.

[03:31](javascript:;)For the supply chain app, external access will be achieved via regional HTTPS load balancer.

[03:40](javascript:;)Three ways you can interact with Google Cloud to work with and deploy services are via Cloud Console, the command line, and programmatically.

[03:50](javascript:;)Let's look at these in a little more detail.

[03:52](javascript:;)You want to implement a compute instance for the Cymbal Superstore development team to start developing code.

[03:59](javascript:;)One of the ways you can do this is view the Google Cloud Console.

[04:04](javascript:;)The screenshot shows some of the settings you need to specify as you create this instance.

[04:10](javascript:;)The name of the instance, the region and zone where the instance resides, the machine configuration, the

[04:17](javascript:;)boot disk, network settings and any other persistent disks that you're going to attach to this virtual machine.

[04:25](javascript:;)Cymbal Superstore's supply chain app needs this Cloud SQL back end.

[04:30](javascript:;)Here is an example of how you would do this via the command line interface.

[04:36](javascript:;)Notice the parameters required include the Name, Resources and Region.

[04:43](javascript:;)Remember, you can access the CLI by loading the Google Cloud SDK on your local machine.

[04:48](javascript:;)You can also use Cloud Shell, a cloud based terminal with a G-Cloud CLI already installed on it.

[04:57](javascript:;)The transportation management system is using Cloud Functions.

[05:01](javascript:;)Cloud Functions give you the option of deploying your function code from the local directory where it resides.

[05:08](javascript:;)Here is an example of the command to deploy a cloud function with a Pub/Sub trigger from a directly on

[05:13](javascript:;)your local machine; trans\_mg\_function is going to be the name of the deployed function based on the logic in the directory.

[05:26](javascript:;)The runtime parameter specifies the Python interpreter you want to use as you parse the function.

[05:33](javascript:;)The trigger topic parameter is the Pub/Sub topic you want to monitor.

[05:39](javascript:;)The data sent to your function includes the Pub/Sub event data and metadata.

Introduction: Diagnostic Questions-

[00:00](javascript:;)>> Now that you have some context for the objectives in this section, it's time to take a self-assessment focused on deploying and implementing Google Cloud solutions.

[00:11](javascript:;)Please take 15 minutes to complete the diagnostic questions for this section.

Diagnostic Questions-

What action does the terraform apply command perform?

Sets up resources requested in the terraform config file.

You require a Cloud Storage bucket serving users in New York City. There is a need for geo-redundancy. You do not plan on using ACLs. What CLI command do you use?

Run a gsutil mb command specifying a dual-region bucket and an option to turn ACL evaluation off.

Correct! NAM4 implements a dual-region bucket with us-east1 and us-central1 as the configured regions.

Cymbal Superstore’s marketing department needs to load some slowly changing data into BigQuery. The data arrives hourly in a Cloud Storage bucket. You want to minimize cost and implement this in the fewest steps. What should you do?  
Use the BigQuery data transfer service to schedule a transfer between your bucket and BigQuery.

Correct! BigQuery transfer service is the simplest process to set up transfers between Cloud Storage and BigQuery. It is encompassed by one command. It is also free.

Which Virtual Private Cloud (VPC) network type allows you to fully control IP ranges and the definition of regional subnets?

Correct! A custom mode network gives you control over regions that you place your subnets in and lets you specify IP ranges for them as well.

Cymbal Superstore’s sales department has a medium-sized MySQL database. This database includes user-defined functions and is used internally by the marketing department at Cymbal Superstore HQ. The sales department asks you to migrate the database to Google Cloud in the most timely and economical way. What should you do?

Configure a Compute Engine VM with an N2 machine type, install MySQL, and restore your data to the new instance.

Correct! N2 is a balanced machine type, which is recommended for medium-large databases.

The development team for the supply chain project is ready to start building their new cloud app using a small Kubernetes cluster for the pilot. The cluster should only be available to team members and does not need to be highly available. The developers also need the ability to change the cluster architecture as they deploy new capabilities. How would you implement this?

Implement a private standard zonal cluster in us-central1-a with a default pool and an Ubuntu image.

Correct! Standard clusters can be zonal. The default pool provides nodes used by the cluster.

Cymbal Superstore asks you to implement Cloud SQL as a database backend to their supply chain application. You want to configure automatic failover in case of a zone outage. You decide to use the gcloud sql instances create command set to accomplish this. Which gcloud command line argument is required to configure the stated failover capability as you create the required instances?  
--availability-type

Correct! This option allows you to specify zonal or regional availability, with regional providing automatic failover to a standby node in another region.

You need to analyze and act on files being added to a Cloud Storage bucket. Your programming team is proficient in Python. The analysis you need to do takes at most 5 minutes. You implement a Cloud Function to accomplish your processing and specify a trigger resource pointing to your bucket. How should you configure the --trigger-event parameter using gcloud?  
--trigger-event google.storage.object.finalize

Correct! Finalize event trigger when a write to Cloud Storage is complete.

The backend of Cymbal Superstore’s e-commerce system consists of managed instance groups. You need to update the operating system of the instances in an automated way using minimal resources. What do you do?

Create a new instance template, then click Update VMs. Set the update type to PROACTIVE. Click Start.

Correct! This institutes a rolling update where the surge is set to 1 automatically, which minimizes resources as requested.

You need to quickly deploy a containerized web application on Google Cloud. You know the services you want to be exposed. You do not want to manage infrastructure. You only want to pay when requests are being handled and need support for custom packages. What technology meets these needs?

Correct! Cloud Run is serverless, exposes your services as an endpoint, and abstracts all infrastructure.

Your Study Plan-

[00:00](javascript:;)>> There's much to cover in this section of the Exam Guide.

[00:04](javascript:;)Just as there's much for an associate cloud engineer to do when deploying and implementing cloud solutions.

[00:11](javascript:;)Let's review the diagnostic questions to help you target your study time, to focus on the areas you need to develop your skills in the most.

[00:20](javascript:;)We'll approach this review by looking at the objectives of this exam section, and the questions you answered about each objective.

[00:28](javascript:;)We'll introduce an objective, briefly review the answers to the related questions, then talk about

[00:34](javascript:;)where you can find out more in the learning resources, or in the Google Cloud documentation.

[00:41](javascript:;)As we go through each section objective, use the page in your workbook to mark the

[00:46](javascript:;)specific documentation, courses or modules and quests that you'll want to emphasize in your study plan.

[00:54](javascript:;)As you can see, there are multiple objectives in this section that have many related tasks.

[00:59](javascript:;)So you will probably need to plan for more study time in this section.

Knowledge Check-

Which data storage service is a unique globally available, horizontally scalable database with relational semantics?

Correct! Cloud Spanner is Google Çloud’s globally available, horizontally scalable relational database.

Which services are based on logic implemented in containers? (Pick two).

Correct! Google Kubernetes Engine is Google Cloud’s managed Kubernetes environment that lets you deploy containerized apps via pods, deployments, and services you specify.

Correct! Cloud Run is a serverless offering that runs your containerized code when monitored events take place.

ENSURING SUCCESSFUL OPERATION OF A CLOUD SOLUTION:

Module Overview-

[00:00](javascript:;)Person: Welcome to Module 4: Ensuring Successful Operation of a Cloud Solution.

[00:05](javascript:;)In this module, you'll explore the scope of ensuring successful cloud operations.

[00:10](javascript:;)This involves managing the compute, storage and networking resources as well as monitoring and logging tasks.

[00:16](javascript:;)These areas correspond to the fourth section of the Associate Cloud Engineer Exam Guide.

[00:22](javascript:;)We'll start by discussing your role as an Associate Cloud Engineer in managing Cymbal Superstore's cloud solutions.

[00:28](javascript:;)Next, you'll assess your skills in this area through 10 diagnostic questions.

[00:31](javascript:;)When we review the questions, identify the resources you'll want to include in your study plan.

Managing Cymbal Superstore’s cloud solutions-

[00:00](javascript:;)Person: Now that Cymbal Superstore's cloud solutions have been deployed and implemented, your role as an Associate Cloud Engineer shifts focus to maintaining successful operations.

[00:10](javascript:;)Let's explore what that means.

[00:12](javascript:;)In deploying and implementing Cymbal Superstore's cloud architecture, you needed to know how to work with various compute, storage and networking resources on Google Cloud.

[00:23](javascript:;)To ensure successful operations, an Associate Cloud Engineer needs the knowledge and skills to manage the resources used in an organization's cloud solutions.

[00:33](javascript:;)You also need to be able to use Google Cloud's operations suite for monitoring and logging.

[00:38](javascript:;)Cymbal Superstore's supply chain management app is made of resources implemented close to their headquarters in Minneapolis, Minnesota.

[00:46](javascript:;)It is architected using Compute Engine.

[00:49](javascript:;)Manage instance groups let their application scale automatically and remain available across zonal outages.

[00:56](javascript:;)Sometimes the instance template that the group is based on might need to be changed.

[01:01](javascript:;)Some reasons you might want to do this include the following.

[01:05](javascript:;)Upgrading the operating system of your instances, conducting A/B or canary testing of capability upgrades, changing the disk type or attach disks attached to your instances.

[01:17](javascript:;)Once you do update the template, you need to ensure the change is propagated to all the VM instances in the group.

[01:24](javascript:;)Cymbal Superstore's e-commerce app is architected using containers deployed to GKE pods.

[01:30](javascript:;)As an Associate Cloud Engineer on the e-commerce team, you might be asked to configure and monitor external connectivity.

[01:38](javascript:;)An external HTTPS load balancer is a solution that advertises a single global IP, provides

[01:43](javascript:;)content close to your end user and forwards content to back ends that are available globally.

[01:51](javascript:;)Cymbal Superstore's transportation management app uses cloud functions to monitor incoming sensor data and implements a cloud data flow pipeline that uses a sync to write data to Bigtable.

[02:03](javascript:;)As an Associate Cloud Engineer it is common to provide information about the sources and syncs required by a pipeline to the data engineer responsible for developing it.

[02:14](javascript:;)You also need to know how to monitor your incoming data stream and manage your cloud function instances.

[02:20](javascript:;)Let's specifically discuss how you might set up resources required to query that data on a regular basis.

[02:28](javascript:;)What would this involve?

[02:30](javascript:;)Let's think about Cymbal Superstore's transportation management app.

[02:34](javascript:;)You can use BigQuery SQL to query your Bigtable data by defining a permanent external table using the Google Cloud Console or the bq command line tool.

[02:44](javascript:;)You do this by creating a table definition file that includes the URI for the table in Bigtable and information about the column families and columns defined in the table.

[02:55](javascript:;)The entries for the table definition are written in JSON.

[02:59](javascript:;)Next, you create an external table reference with a bq mk command.

[03:03](javascript:;)On the screen, you'll see an example of what that might look like for our Cymbal Superstore example.

[03:09](javascript:;)Then, you can submit the query using BigQuery SQL.

Introduction: Diagnostic Questions-

[00:00](javascript:;)>> Now it's your turn to assess your experience and skills related to this section with some diagnostic questions.

[00:06](javascript:;)Remember, the purpose of these questions is to help you better understand what is involved in this

[00:11](javascript:;)section of the Exam Guide, and identify which areas you'll want to focus on in your study plan.

Diagnostic Questions-

You have a Cloud Run service with a database backend. You want to limit the number of connections to your database. What should you do?

Set Max instances.

Correct! Max instances control costs, keeping you from starting too many instances by limiting your number of connections to a backing service.

You want to implement a lifecycle rule that changes your storage type from standard to nearline after a specific date. What conditions should you use? (Pick two).

Correct! MatchesStorageClass is required to look for objects with a standard storage type.

Correct! CreatedBefore lets you specify a date.

Cymbal Superstore’s supply chain management system has been deployed and is working well. You are tasked with monitoring the system’s resources so you can react quickly to any problems. You want to ensure the CPU usage of each of your Compute Engine instances in us-central1 remains below 60%. You want an incident created if it exceeds this value for 5 minutes. You need to configure the proper alerting policy for this scenario. What should you do?

Choose resource type of VM instance and metric of CPU utilization, condition trigger if any time series violates, condition is above, threshold is .60 for 5 minutes.

What is the declarative way to initialize and update Kubernetes objects?

Correct! kubectl apply creates and updates Kubernetes objects in a declarative way from manifest files.

You have a scheduled snapshot you are trying to delete, but the operation returns an error. What should you do to resolve this problem?  
Detach the snapshot schedule before deleting it.

Correct! You can’t delete a snapshot schedule that is still attached to a persistent disk.

Cymbal Superstore has a subnetwork called mysubnet with an IP range of 10.1.2.0/24. You need to expand this subnet to include enough IP addresses for at most 2000 users or devices. What should you do?

gcloud compute networks subnets expand-ip-range mysubnet --region us-central1 --prefix-length 21

Correct! This command gives a total of 2046 addresses available and meets the requirement.

You want to view a description of your available snapshots using the command line interface (CLI). What gcloud command should you use?

gcloud compute snapshots list

Correct! gcloud commands are built with groups and subgroups, followed by a command, which is a verb. In this example, Compute is the Group, snapshots is the subgroup, and list is the command.

Cymbal Superstore’s GKE cluster requires an internal http(s) load balancer. You are creating the configuration files required for this resource. What is the proper setting for this scenario?

Annotate your service object with a “neg" reference.

Correct! This is correct because an internal http(s) load balancer can only use NEGs.

Which of the following tasks are part of the process when configuring a managed instance group? (Pick two).

Correct! Health checks are part of your managed instance group configuration.

Correct! Number of instances is part of your managed instance group configuration.

What Kubernetes object provides access to logic running in your cluster via endpoints that you define?

Services

Correct! Service endpoints are defined by pods with labels that match those specified in the service configuration file. Services then specify how those pods are exposed.

Your Study Plan-

[00:00](javascript:;)>> What areas do you need to develop your skills in order to manage the different aspects of a Google Cloud solution?

[00:06](javascript:;)This is another important area for an associate cloud engineer, and where you'll likely spend much of your time on the job.

[00:13](javascript:;)Let's review the diagnostic questions to help you target your study time to focus on the areas where you need to develop your skills.

[00:21](javascript:;)We'll approach this review by looking at the objectives of this exam section, and the questions you just answered about each one.

[00:28](javascript:;)We'll introduce an objective, briefly review the answers to the related questions, then talk about where you can find out more in the learning resources and/or in Google Cloud documentation.

[00:40](javascript:;)As we go through each section objective, use the page in your workbook to mark

[00:44](javascript:;)the specific documentation, courses and modules, and quests you'll want to emphasize in your study plan.

[00:52](javascript:;)Just like with the previous section, there are multiple objectives in this section that have many related tasks.

[00:58](javascript:;)So you will probably need to plan for more study time.

Knowledge Check-

What GKE object implements an http(s) load balancer?

Ingress

Correct! Ingress objects implement an http(s) load balancer based on upstream services you specify in your configuration file.

Which Cloud Run autoscaling setting should you set if you want to limit cost?

Max instances

Correct! Max instances directly affects cost by limiting the maximum amount of container instances deployed.

CONFIGURING ACCESS AND SECURITY:

Module Overview:

[00:00](javascript:;)Person: Welcome to Module 5: Configuring Access and Security.

[00:06](javascript:;)In this module, you'll explore the scope of configuring access and security.

[00:11](javascript:;)This involves managing IAM as well as service accounts for cloud solutions.

[00:17](javascript:;)These areas correspond to the fifth and last section of the Associate Cloud Engineer Exam Guide.

[00:24](javascript:;)We'll start by discussing your role as an Associate Cloud Engineer in managing access for Cymbal Superstore's cloud solutions.

[00:33](javascript:;)Next, you'll assess your skills in this section of the exam guide through seven diagnostic questions.

[00:39](javascript:;)There are fewer questions for this module.

[00:42](javascript:;)While this is an important area for you to understand as an Associate Cloud Engineer, it is narrower in scope.

[00:51](javascript:;)When we review the questions, identify the resources you'll want to include in your study plan.

Managing Access for Cymbal Superstore’s cloud solutions:

[00:00](javascript:;)Person: As Cymbal Superstore uses its application on Google Cloud, an Associate Cloud Engineer plays an ongoing role in configuring and managing IAM access and service accounts.

[00:12](javascript:;)Let's explore some examples of how you might do this at Cymbal Superstore.

[00:17](javascript:;)To successfully perform the Associate Cloud Engineer role at Cymbal Superstore, you need to be able to manage Identity and Access Management, or IAM, in Google Cloud.

[00:29](javascript:;)We talked about the basics of IAM in the first module from the perspective of setting up cloud projects and accounts.

[00:36](javascript:;)Here, you'll consider skills involved in managing access.

[00:41](javascript:;)You'll also need to be familiar with service accounts and recommended practices to manage them in Google Cloud.

[00:48](javascript:;)You'll also need to know how to view audit logs when required.

[00:52](javascript:;)To give you a better idea of what configuring access and security involves in practice, let's explore an example of where you might use a service account at Cymbal Superstore.

[01:05](javascript:;)Cymbal Superstore's supply chain app is built on a lamp stack using Google Compute Engineer virtual machine instances.

[01:13](javascript:;)It uses Cloud SQL as a backing data store.

[01:17](javascript:;)The app needs to talk to Cloud SQL to update inventory levels.

[01:21](javascript:;)It does this through a service account attached to the virtual machine that it runs on.

[01:28](javascript:;)Service accounts are designed to enable machine-to-machine communication for just this purpose.

[01:34](javascript:;)The first step in setting up a service account for Cymbal Superstore's supply chain app is to create the service account.

[01:42](javascript:;)Next, you assign permissions to the service account you just created.

[01:47](javascript:;)Finally, you attach that service account to a Compute Engine virtual machine.

[01:53](javascript:;)Attaching a service account allows the virtual machine and all the apps running on it to use the permissions assigned to the service account.

[02:01](javascript:;)Let's look at these steps in more detail.

[02:04](javascript:;)Go to the project you want to add the service account to.

[02:07](javascript:;)Service accounts are both identities and managed resources in Google Cloud.

[02:13](javascript:;)Select the service account link in the IAM menu of your project, then select create service account.

[02:21](javascript:;)In the dialog that comes up, name your service account and note the email address associated with it.

[02:28](javascript:;)You can also provide a description of what this service account does.

[02:32](javascript:;)Once you select create, your new service account will be added to the list of all your service accounts.

[02:39](javascript:;)Select the three ellipsis under actions for a list of all the actions you can perform on your new service account.

[02:48](javascript:;)Next, we'll use one of these choices to manage permissions for the service account.

[02:54](javascript:;)Select manage permissions under the actions dialog in the service account list.

[03:00](javascript:;)A new menu let's you pick your service account and add permissions to it.

[03:06](javascript:;)Copy your service account email address identifier.

[03:10](javascript:;)Search or browse the permissions to find the ones you need to add.

[03:15](javascript:;)In our example, we'll give our service account permissions as a Cloud SQL instance user.

[03:23](javascript:;)Finally, when you add your virtual machine instance, you have a chance to add the service account to it under the identity and API access section.

[03:32](javascript:;)This covers authorization.

[03:36](javascript:;)Authentication is another important aspect of both user accounts and service accounts that you should be familiar with as an Associate Cloud Engineer.

Introduction: Diagnostic Questions:

[00:00](javascript:;)>> Now that you have some context for the objectives in this exam section, it's time to take a self-assessment focused on configuring access and security for Google Cloud solutions.

[00:11](javascript:;)Please take 15 minutes to complete the diagnostic questions for this section.

Diagnostic Questions:

Cymbal Superstore is implementing a mobile app for end users to track deliveries that are en route to them. The app needs to access data about truck location from Pub/Sub using Google recommended practices. What kind of credentials should you use?

Correct! Service account keys are used for accessing private data such as your Pub/Sub truck information from an external environment such as a mobile app running on a phone.

Which of the scenarios below is an example of a situation where you should use a service account?

For individual GKE pods

Correct! When configuring access for GKE, you set up dedicated service accounts for each pod. You then use workload identity to map them to dedicated Kubernetes service accounts.

Outline where Cloud Audit logs can be accessed: in the logging tab of the operations interface Link: <https://cloud.google.com/storage/docs/audit-logging> You are configuring audit logging for Cloud Storage. You want to know when objects are added to a bucket. Which type of audit log entry should you monitor?

Correct! DATA\_WRITE log entries include information about when objects are created or deleted.

You have a custom role implemented for administration of the dev/test environment for Cymbal Superstore’s transportation management application. You are developing a pilot to use Cloud Run instead of Cloud Functions. You want to ensure your administrators have the correct access to the new resources. What should you do?

Make the change to the custom role locally and run an update on the custom role

Correct! There is a recommended process to update an existing custom role. You get the current policy, update it locally, and write the updated policy back into Google Cloud. The gcloud commands used in this process include the get and update policy subcommands.

You are trying to assign roles to the dev and prod projects of Cymbal Superstore’s e-commerce app but are receiving an error when you try to run set-iam policy. The projects are organized into an ecommerce folder in the Cymbal Superstore organizational hierarchy. You want to follow best practices for the permissions you need while respecting the practice of least privilege. What should you do?

Ask your administrator for the roles/resourcemanager.folderIamAdmin for the ecommerce folder

Correct! This choice gives you the required permissions while minimizing the number of individual resources you have to set permissions for.

You need to configure access to Cloud Spanner from the GKE cluster that is supporting Cymbal Superstore’s ecommerce microservices application. You want to specify an account type to set the proper permissions. What should you do?

Assign permissions through service account referenced by the application

Correct! A service account uses an account identity and an access key. It is used by applications to connect to services.

Which Cloud Audit log is disabled by default with a few exceptions?

Correct! Data Access audit logs are disabled by default except for BigQuery.

Your Study Plan-

[00:00](javascript:;)>> What areas do you need to develop your skills in, in order to successfully manage access for Google Cloud solutions?

[00:07](javascript:;)Recall that this exam section is narrower in scope, but nevertheless important for the role of an Associate Cloud Engineer.

[00:16](javascript:;)Let's review the diagnostic questions to help you target your study time to focus on the areas where you need to develop your skills.

[00:24](javascript:;)We'll approach this review by looking at the objectives of this exam section and the questions you just answered about each objective.

[00:32](javascript:;)We'll introduce an objective, briefly review the answers to the related questions, and then talk about

[00:37](javascript:;)where you can find out more either in the learning resources or in the Google Cloud documentation.

[00:44](javascript:;)As we go through each section objective, use the page in your workbook to mark the

[00:50](javascript:;)specific documentation, courses or modules, and the quest that you'll want to emphasize in your study plan.

[00:57](javascript:;)There are fewer objectives and tasks involved with this section, but don't forget to plan for them in your study time.

Knowledge Check-

What kind of account is meant for machine-to-machine communication in Google Cloud?

Correct! Service accounts are designed to provide permissions for machine-to machine or service-to-service communications in Google Cloud.

You are authenticating an application to service APIs. Both resources are internal to the Google Cloud environment. What type of credentials should you use?

Temporary credentials

Correct! You should use temporary credentials of the service account provided by the environment.

YOUR NEXT STEPS:

[00:01](javascript:;)person: Welcome to Module 6: Your Next Steps.

[00:03](javascript:;)In this module, you'll focus on creating your individualized study plan.

[00:08](javascript:;)In this module, you'll use the notes you've been taking throughout this course to put together a study plan for each week in your Associate Cloud Engineer journey.

[00:17](javascript:;)Now that you've explored all five sections of the exam guide, consider what you've learned about your knowledge and skills through the diagnostic questions in this course.

[00:26](javascript:;)You should have a better understanding of what areas you need to focus on and what resources are available.

[00:32](javascript:;)Think about the answers to these questions: When will you take the exam?

[00:37](javascript:;)How many weeks does that give you to prepare?

[00:40](javascript:;)How many hours can you realistically spend preparing for the exam each week?

[00:45](javascript:;)How many total hours will you prepare?

[00:48](javascript:;)Be sure to leave enough time at the end of your plan to retake the diagnostic

[00:50](javascript:;)questions and the sample questions and fill in any gaps in your knowledge that may remain.

[00:57](javascript:;)Take a few minutes to think about how much time you will allocate to preparing for the exam and note your answers in the workbook.

[01:04](javascript:;)The number of weeks in your preparation journey will depend on a variety of factors, such as your

[01:08](javascript:;)prior experience with Google Cloud and how much time you have available to dedicate to studying each week.

[01:15](javascript:;)You might choose to focus on specific courses or skill badges each week, such as in this

[01:19](javascript:;)sample study plan or instead focus your weekly study on a specific topic, such as configuring VPCs.

[01:27](javascript:;)Once you have a high level idea of how many weeks you have to study and how you

[01:30](javascript:;)want to determine your weekly focus, you'll want to build out a plan with weekly goals and study activities.

[01:37](javascript:;)Use the template in your workbook to plan your study goals for each week.

[01:40](javascript:;)Consider: What exam guide sections or topic areas will you focus on?

[01:46](javascript:;)What courses or specific modules will help you learn more?

[01:51](javascript:;)What skill badges or labs will you work on for hands-on practice?

[01:56](javascript:;)What documentation links will you review?

[01:59](javascript:;)What additional resources will you use, such as sample questions?

[02:03](javascript:;)You may do some or all of these study activities each week.

[02:07](javascript:;)Let's review an example.

[02:08](javascript:;)If you've identified configuring access using IAM as a particular area you need to study, you might choose to structure your

[02:14](javascript:;)study for a week to include targeted modules from the on-demand training, a related skill badge for hands-on practice and documentation.

[02:24](javascript:;)Alternately, you might choose one week to complete an entire course and another week to focus on a skill badge.

[02:31](javascript:;)You can determine the approach that fits your existing skillset.

[02:34](javascript:;)Find the weekly study template at the end of your workbook.

[02:38](javascript:;)Duplicate the weekly template for the number of weeks in your individual preparation journey.

[02:43](javascript:;)Remember, you may need to adjust your plans based on the areas where you need to learn more.

[02:49](javascript:;)For more information about the resources we've discussed in this course, refer to your notes and the student copy of the slides.

[02:56](javascript:;)To register for the exam, follow the link on the ACE certification information page using the URL found on the slide.

[03:04](javascript:;)Thank you for attending the preparing for your Associate Cloud Engineer journey.

[03:08](javascript:;)Good luck as you begin your journey to study for the Associate Cloud Engineer certification.

====================================================================

GOOGLE CLOUD FUNDAMENTALS: CORE INFRASTRUCTURE

Course Introduction-

[00:00](javascript:;)Hi and welcome to the Google Cloud Fundamentals: Core Infrastructure training course.

[00:04](javascript:;)The goal of this course is to provide you with an overview of Google Cloud.

[00:10](javascript:;)Google Cloud offerings can be broadly categorised as Compute storage Big data Machine learning and application services for web, mobile, analytics, and back-end solutions.

[00:23](javascript:;)Through a combination of videos, quizzes, and hands-on labs, you’ll learn the value of Google Cloud and how cloud solutions factor into business strategies.

[00:33](javascript:;)The intended target audience of today’s course consists of solutions developers, systems operations professionals, and solution architects planning to deploy applications and create application environments on Google Cloud.

[00:45](javascript:;)The course will also be useful for business decision makers evaluating Google Cloud.

[00:51](javascript:;)While you should all be happy to hear that we’ll be finding out about services and concepts that are specific to Google Cloud in today's

[00:56](javascript:;)course, do keep in mind that, as a ‘fundamentals’ level course, some content will be geared towards learners who are entirely new to cloud technologies.

[01:05](javascript:;)The course has no prerequisites, Although it’s helpful be familiar with application development, Linux operating systems, systems operations, and data analytics/machine learning to best understand the technologies covered.

[01:21](javascript:;)There are five key learning objectives that we’re hoping to achieve.

[01:23](javascript:;)By the end of this course you should be able to: Identify the purpose and value of Google Cloud products and services.

[01:32](javascript:;)Choose among and use application deployment environments on Google Cloud: App Engine, Google Kubernetes Engine, and Compute Engine.

[01:40](javascript:;)Choose among and use Google Cloud storage options: Cloud Storage, Cloud SQL, Cloud Bigtable, and Firestore.

[01:50](javascript:;)Interact with Google Cloud services.

[01:53](javascript:;)Describe ways in which customers have used Google Cloud.

[01:56](javascript:;)OK, all set?

[01:58](javascript:;)Let’s begin!

INTRODUCING GOOGLE CLOUD:

Cloud Computing Overview-

[00:00](javascript:;)Let’s start at the beginning with an overview of cloud computing.

[00:04](javascript:;)The cloud is a hot topic these days, but what exactly is it?

[00:10](javascript:;)The US National Institute of Standards and Technology created the term cloud computing, although there is nothing US-specific about it.

[00:17](javascript:;)Cloud computing is a way of using information technology (IT) that has these five equally important traits First, customers get computing resources that are on-demand and self-service.

[00:31](javascript:;)Through a web interface, users get the processing power, storage, and network they need with no need for human intervention.

[00:40](javascript:;)Second, customers get access to those resources over the internet, from anywhere they have a connection.

[00:46](javascript:;)Third, the cloud provider has a big pool of those resources and allocates them to users out of that pool.

[00:53](javascript:;)That allows the provider to buy in bulk and pass the savings on to the customers.

[00:57](javascript:;)Customers don't have to know or care about the exact physical location of those resources.

[01:04](javascript:;)Fourth, the resources are elastic–which means they’re flexible, so customers can be.

[01:10](javascript:;)If they need more resources they can get more, and quickly.

[01:15](javascript:;)If they need less, they can scale back.

[01:17](javascript:;)And finally, customers pay only for what they use, or reserve as they go.

[01:23](javascript:;)If they stop using resources, they stop paying.

[01:26](javascript:;)That's it.

[01:27](javascript:;)That's the definition of cloud.

[01:30](javascript:;)But why is the cloud model so compelling nowadays?

[01:33](javascript:;)To understand why, we need to look at some history.

[01:37](javascript:;)The trend towards cloud computing started with a first wave known as colocation Colocation gave

[01:43](javascript:;)users the financial efficiency of renting physical space, instead of investing in data center real estate.

[01:49](javascript:;)Virtualized data centers of today, which is the second wave, share similarities with the private data centers and colocation facilities of decades past.

[02:00](javascript:;)The components of virtualized data centers match the physical building blocks of hosted computing—servers, CPUs, disks, load balancers, and so on—but now they’re virtual devices.

[02:13](javascript:;)With virtualization, enterprises still maintain the infrastructure; but it also remains a user-controlled and user-configured environment.

[02:22](javascript:;)Several years ago, Google realized that its business couldn’t move fast enough within the confines of the virtualization model.

[02:30](javascript:;)So Google switched to a container-based architecture—a fully automated, elastic third-wave cloud that consists of a combination of automated services and scalable data.

[02:42](javascript:;)Services automatically provision and configure the infrastructure used to run applications.

[02:47](javascript:;)Today, Google Cloud makes this third-wave cloud available to Google customers.

[02:54](javascript:;)Google believes that, in the future, every company—regardless of size or industry— will differentiate itself from its competitors through technology.

[03:03](javascript:;)Increasingly, that technology will be in the form of software.

[03:07](javascript:;)Great software is based on high-quality data.

[03:12](javascript:;)This means that every company is, or will eventually become, a data company.

IaaS and PaaS-

[00:00](javascript:;)The move to virtualized data centers introduced customers to two new types of offerings: infrastructure as a service, commonly referred to as IaaS, and platform as a service, or PaaS.

[00:13](javascript:;)IaaS offerings provide raw compute, storage, and network capabilities, organized virtually into resources that are similar to physical data centers.

[00:22](javascript:;)PaaS offerings, in contrast, bind code to libraries that provide access to the infrastructure application needs.

[00:31](javascript:;)This allows more resources to be focused on application logic.

[00:36](javascript:;)In the IaaS model, customers pay for the resources they allocate ahead of time; in the PaaS model, customers pay for the resources they actually use.

[00:46](javascript:;)As cloud computing has evolved, the momentum has shifted toward managed infrastructure and managed services.

[00:54](javascript:;)Leveraging managed resources and services allows companies to concentrate more on their business goals and spend less time and money on creating and maintaining their technical infrastructure.

[01:05](javascript:;)It allows companies to deliver products and services to their customers more quickly and reliably.

[01:11](javascript:;)Serverless is yet another step in the evolution of cloud computing.

[01:16](javascript:;)It allows developers to concentrate on their code, rather than on server configuration, by eliminating the need for any infrastructure management.

[01:26](javascript:;)Serverless technologies offered by Google include Cloud Functions, which manages event-driven code as a pay-as-you-go service,

[01:31](javascript:;)and Cloud Run, which allows customers to deploy their containerized microservices based application in a fully-managed environment.

[01:40](javascript:;)While it’s outside the scope of this course, you might have heard about software as

[01:45](javascript:;)a service, SaaS, and wondered what it is and how it fits into the Cloud ecosphere.

[01:52](javascript:;)Software as a Service applications are not installed on your local computer.

[01:56](javascript:;)Instead, they run in the cloud as a service and are consumed directly over the internet by end users Popular

[02:03](javascript:;)Google applications such as Gmail, Docs, and Drive, that are a part of Google Workspace, are all examples of SaaS.

The Google Cloud Network-

[00:00](javascript:;)Google Cloud runs on Google’s own global network.

[00:03](javascript:;)It’s the largest network of its kind, and Google has invested billions of dollars over many years to build it.

[00:11](javascript:;)This network is designed to give customers the highest possible throughput and lowest possible latencies for their applications by leveraging more than 100 content caching nodes worldwide.

[00:23](javascript:;)These are locations where high demand content is cached for quicker access, allowing applications to respond to user requests from the location that will provide the quickest response time.

[00:34](javascript:;)Google Cloud’s infrastructure is based in five major geographic locations: North America, South America, Europe, Asia, and Australia.

[00:46](javascript:;)Having multiple locations is important because choosing where to locate applications affects qualities like availability, durability, and latency, the

[00:55](javascript:;)latter of which measures the time a packet of information takes to travel from its source to its destination.

[01:04](javascript:;)Each of these locations is divided into several different regions and zones.

[01:09](javascript:;)Regions represent independent geographic areas and are composed of zones.

[01:14](javascript:;)For example, London, or europe-west2, is a region that currently comprises three different zones.

[01:22](javascript:;)A zone is an area where Google Cloud resources are deployed.

[01:27](javascript:;)For example, if you launch a virtual machine using Compute Engine it will run in the zone that you specify to ensure resource redundancy.

[01:37](javascript:;)You can also run resources in different regions.

[01:40](javascript:;)This is useful for bringing applications closer to users around the world, and also for protection in case there are issues with an entire region, such as a natural disaster.

[01:50](javascript:;)Some of Google Cloud’s services support placing resources in what we call a multi-region.

[01:56](javascript:;)For example, Cloud Spanner multi-region configurations allow you to replicate the database's data not just

[02:02](javascript:;)in multiple zones, but in multiple zones across multiple regions, as defined by the instance configuration.

[02:10](javascript:;)These additional replicas enable you to read data with low latency from multiple locations close to or within the regions in the configuration, like The Netherlands, and Belgium.

[02:21](javascript:;)Google Cloud currently supports 103 zones in 34 regions, although this number is increasing all the time.

[02:29](javascript:;)You can find the most up-to-date numbers at cloud.google.com/about/locations.

Environmental Impact-

[00:00](javascript:;)The virtual world, which includes Google Cloud’s network, is built on physical infrastructure, and all those racks of humming servers use huge amounts of energy.

[00:09](javascript:;)Altogether, existing data centers use roughly 2% of the world’s electricity.

[00:16](javascript:;)With this in mind, Google works to make their data centers run as efficiently as possible.

[00:22](javascript:;)Just like our customers, Google is trying to do the right things for the planet.

[00:26](javascript:;)We understand that Google Cloud customers have environmental goals of their own, and running their workloads on Google Cloud can be a part of meeting those goals.

[00:34](javascript:;)Therefore, it’s useful to note that Google's data centers were the first to achieve ISO 14001 certification, which is a

[00:42](javascript:;)standard that maps out a framework for an organization to enhance its environmental performance through improving resource efficiency and reducing waste.

[00:53](javascript:;)As an example of how this is being done, here’s Google’s data center in Hamina, Finland.

[00:59](javascript:;)This facility is one of the most advanced and efficient data centers in the Google fleet.

[01:03](javascript:;)Its cooling system, which uses sea water from the Bay of Finland, reduces energy use and is the first of its kind anywhere in the world.

[01:13](javascript:;)In our founding decade, Google became the first major company to be carbon neutral.

[01:18](javascript:;)In our second decade, we were the first company to achieve 100% renewable energy.

[01:24](javascript:;)By 2030, we aim to be the first major company to operate completely carbon free.

Security-

[00:00](javascript:;)Nine of Google’s services have more than one billion users each, and so you can be assured that security is always on the minds of Google's employees.

[00:08](javascript:;)Design for security is prevalent throughout the infrastructure that Google Cloud and Google services run on.

[00:15](javascript:;)Let's talk about a few ways Google works to keep customers' data safe.

[00:20](javascript:;)The security infrastructure can be explained in progressive layers, starting from the physical security of our data centers, continuing on to how the

[00:27](javascript:;)hardware and software that underlie the infrastructure are secured, and finally, describing the technical constraints and processes in place to support operational security.

[00:39](javascript:;)We begin with the Hardware infrastructure layer which comprises three key security features: The first is hardware design and provenance.

[00:49](javascript:;)Both the server boards and the networking equipment in Google data centers are custom-designed by Google.

[00:56](javascript:;)Google also designs custom chips, including a hardware security chip that's currently being deployed on both servers and peripherals.

[01:05](javascript:;)The next feature is a secure boot stack.

[01:09](javascript:;)Google server machines use a variety of technologies to ensure that they are booting the correct

[01:14](javascript:;)software stack, such as cryptographic signatures over the BIOS, bootloader, kernel, and base operating system image.

[01:23](javascript:;)This layer's final feature is premises security.

[01:27](javascript:;)Google designs and builds its own data centers, which incorporate multiple layers of physical security protections.

[01:34](javascript:;)Access to these data centers is limited to only a very small number of Google employees.

[01:40](javascript:;)Google additionally hosts some servers in third-party data centers, where we ensure that there are

[01:44](javascript:;)Google-controlled physical security measures on top of the security layers provided by the data center operator.

[01:51](javascript:;)Next is the Service deployment layer, where the key feature is encryption of inter-service communication.

[01:57](javascript:;)Google’s infrastructure provides cryptographic privacy and integrity for remote procedure call (“RPC”) data on the network.

[02:05](javascript:;)Google’s services communicate with each other using RPC calls.

[02:08](javascript:;)The infrastructure automatically encrypts all infrastructure RPC traffic that goes between data centers.

[02:15](javascript:;)Google has started to deploy hardware cryptographic accelerators that will allow it to extend this default encryption to all infrastructure RPC traffic inside Google data centers.

[02:28](javascript:;)Then we have the User identity layer.

[02:31](javascript:;)Google’s central identity service, which usually manifests to end users as the Google login page, goes beyond asking for a simple username and password.

[02:41](javascript:;)The service also intelligently challenges users for additional information based on risk factors such as

[02:45](javascript:;)whether they have logged in from the same device or a similar location in the past.

[02:51](javascript:;)Users can also employ secondary factors when signing in, including devices based on the Universal 2nd Factor (U2F) open standard.

[03:01](javascript:;)On the Storage services layer we find the encryption at rest security feature.

[03:07](javascript:;)Most applications at Google access physical storage (in other words, “file storage”) indirectly via storage

[03:12](javascript:;)services, and encryption using centrally managed keys is applied at the layer of these storage services.

[03:21](javascript:;)Google also enables hardware encryption support in hard drives and SSDs.

[03:25](javascript:;)The next layer up is the Internet communication layer, and this comprises two key security features.

[03:33](javascript:;)Google services that are being made available on the internet, register themselves with an infrastructure service called the Google Front End, which ensures that all TLS connections

[03:42](javascript:;)are ended using a public-private key pair and an X.509 certificate from a Certified Authority (CA), as well as following best practices such as supporting perfect forward secrecy.

[03:54](javascript:;)The GFE additionally applies protections against Denial of Service attacks.

[04:00](javascript:;)Also provided is Denial of Service (“DoS”) protection.

[04:04](javascript:;)The sheer scale of its infrastructure enables Google to simply absorb many DoS attacks.

[04:09](javascript:;)Google also has multi-tier, multi-layer DoS protections that further reduce the risk of any DoS impact on a service running behind a GFE.

[04:20](javascript:;)The final layer is Google's Operational security layer which provides four key features.

[04:27](javascript:;)First is intrusion detection.

[04:29](javascript:;)Rules and machine intelligence give Google’s operational security teams warnings of possible incidents.

[04:36](javascript:;)Google conducts Red Team exercises to measure and improve the effectiveness of its detection and response mechanisms.

[04:44](javascript:;)Next is reducing insider risk.

[04:46](javascript:;)Google aggressively limits and actively monitors the activities of employees who have been granted administrative access to the infrastructure.

[04:54](javascript:;)Then there’s employee U2F use.

[04:57](javascript:;)To guard against phishing attacks against Google employees, employee accounts require use of U2F-compatible Security Keys.

[05:05](javascript:;)Finally, there are stringent software development practices.

[05:10](javascript:;)Google employs central source control and requires two-party review of new code.

[05:16](javascript:;)Google also provides its developers libraries that prevent them from introducing certain classes of security bugs.

[05:22](javascript:;)Additionally, Google runs a Vulnerability Rewards Program where we pay anyone who is able to discover and inform us of bugs in our infrastructure or applications.

[05:31](javascript:;)You can learn more about Google’s technical-infrastructure security at cloud.google.com/security/security-design

Open Source Ecosystems-

[00:00](javascript:;)Some organizations are afraid to bring their workloads to the cloud because they're afraid they'll get locked into a particular vendor.

[00:06](javascript:;)However, if, for whatever reason, a customer decides that Google is no longer the best provider for their needs, we provide them with the ability to run their applications elsewhere.

[00:17](javascript:;)Google publishes key elements of technology using open source licenses to create ecosystems that provide customers with options other than Google.

[00:26](javascript:;)For example, TensorFlow, an open source software library for machine learning developed inside Google, is at the heart of a strong open source ecosystem.

[00:35](javascript:;)Google provides interoperability at multiple layers of the stack.

[00:39](javascript:;)Kubernetes and Google Kubernetes Engine give customers the ability to mix and match microservices running

[00:45](javascript:;)across different clouds, while Google Cloud’s operations suite lets customers monitor workloads across multiple cloud providers.

Pricing and Billing-

[00:00](javascript:;)To round off this section of the course, let’s take a brief look at Google Cloud’s pricing structure.

[00:05](javascript:;)Google was the first major cloud provider to deliver per-second billing for its infrastructure-as-a-service compute offering, Compute Engine.

[00:13](javascript:;)In addition, per-second billing is now also offered for users of Google Kubernetes Engine (our container infrastructure as a service), Dataproc (which is

[00:22](javascript:;)the equivalent of the big data system Hadoop, but operating as a service), and App Engine flexible environment VMs (a platform as a service).

[00:32](javascript:;)We’ll explore these products and services later in the course.

[00:37](javascript:;)Compute Engine offers automatically applied sustained-use discounts, which are automatic discounts that you get for running a virtual machine instance for a significant portion of the billing month.

[00:48](javascript:;)Specifically, when you run an instance for more than 25% of a month, Compute Engine automatically gives you a discount for every incremental minute you use for that instance.

[00:58](javascript:;)Custom virtual machine types allow Compute Engine virtual machines to be fine-tuned with optimal amounts of

[01:05](javascript:;)vCPU and memory for their applications so that you can tailor your pricing for your workloads.

[01:12](javascript:;)Our online pricing calculator can help estimate your costs.

[01:16](javascript:;)Visit cloud.google.com/products/calculator to try it out.

[01:22](javascript:;)Now, you’re probably thinking, “How can I make sure I don’t accidentally run up a big Google Cloud bill?”

[01:30](javascript:;)We provide a few tools to help.

[01:33](javascript:;)You can define budgets at the billing account level or at the project level.

[01:38](javascript:;)A budget can be a fixed limit, or it can be tied to another metric; for example, a percentage of the previous month’s spend.

[01:46](javascript:;)To be notified when costs approach your budget limit, you can create an alert.

[01:51](javascript:;)For example, with a budget limit of $20,000 and an alert set at 90%, you’ll receive a notification alert when your expenses reach $18,000.

[02:01](javascript:;)Alerts are generally set at 50%, 90% and 100%, but can also be customized.

[02:10](javascript:;)Reports is a visual tool in the Google Cloud Console that allows you to monitor expenditure based on a project or services.

[02:18](javascript:;)Finally, Google Cloud also implements quotas, which are designed to prevent the over-consumption of resources because of

[02:24](javascript:;)an error or a malicious attack, protecting both account owners and the Google Cloud community as a whole.

[02:33](javascript:;)There are two types of quotas: rate quotas and allocation quotas.

[02:38](javascript:;)Both are applied at the project level.

[02:41](javascript:;)Rate quotas reset after a specific time.

[02:44](javascript:;)For example, by default, the GKE service implements a quota of 1,000 calls to its API from each Google Cloud project every 100 seconds.

[02:53](javascript:;)After that 100 seconds, the limit is reset.

[02:59](javascript:;)Allocation quotas govern the number of resources you can have in your projects.

[03:03](javascript:;)For example, by default, each Google Cloud project has a quota allowing it no more than five Virtual Private Cloud networks.

[03:11](javascript:;)Although projects all start with the same quotas, you can change some of them by requesting an increase from Google Cloud Support.

Quiz: Introducing Google Cloud-

What cloud computing service binds application code to libraries that give access to the infrastructure an application needs?

Platform as a service

What cloud computing service provides raw compute, storage, and network resources that are organized similarly to physical data centers?

Infrastructure as a service

Who benefits the most from billing by the second for cloud resources, such as virtual machines?

Customers who create and run many virtual machines

What is the primary benefit to a Google Cloud customer of using resources in several zones within a region?

For improved fault tolerance

Which one of the following statements is true regarding the ability to scale cloud computing resources up and down?

CPU, memory, and storage resources are elastic.

Select two fundamental characteristics of cloud computing from this list.

Customers can scale their resource use up and down.

Resources are available from anywhere over the network.

RESOURCE AND ACCESS IN THE CLOUD:

Google Cloud Resource Hierarchy-

[00:00](javascript:;)In this section of the course we’ll look at the functional structure of Google Cloud.

[00:05](javascript:;)Google Cloud’s resource hierarchy contains four levels, and starting from the bottom up they are: resources, projects, folders, and an organization node.

[00:17](javascript:;)At the first level are resources.

[00:20](javascript:;)These represent virtual machines, Cloud Storage buckets, tables in BigQuery, or anything else in Google Cloud.

[00:29](javascript:;)Resources are organized into projects, which sit on the second level.

[00:34](javascript:;)Projects can be organized into folders, or even subfolders.

[00:37](javascript:;)These sit at the third level.

[00:40](javascript:;)And then at the top level is an organization node, which encompasses all the projects, folders, and resources in your organization.

[00:47](javascript:;)It’s important to understand this resource hierarchy because it directly relates to how policies are managed and applied when you use Google Cloud.

[00:57](javascript:;)Policies can be defined at the project, folder, and organization node levels.

[01:02](javascript:;)Some Google Cloud services allow policies to be applied to individual resources, too.

[01:09](javascript:;)Policies are also inherited downward.

[01:11](javascript:;)This means that if you apply a policy to a folder, it will also apply to all of the projects within that folder.

[01:17](javascript:;)Let’s take a look at the second level of the resource hierarchy, projects, in a little more detail.

[01:24](javascript:;)Projects are the basis for enabling and using Google Cloud services, like managing APIs, enabling billing, adding and removing collaborators, and enabling other Google services.

[01:36](javascript:;)Each project is a separate entity under the organization node, and each resource belongs to exactly one project.

[01:44](javascript:;)Projects can have different owners and users because they’re billed and managed separately.

[01:49](javascript:;)Each Google Cloud project has three identifying attributes: a project ID, a project name, and a project number.

[01:57](javascript:;)The project ID is a globally unique identifier assigned by Google that can’t be changed after creation.

[02:04](javascript:;)They’re what we refer to as being immutable.

[02:08](javascript:;)Project IDs are used in different contexts to inform Google Cloud of the exact project to work with.

[02:15](javascript:;)Project names, however, are user-created.

[02:17](javascript:;)They don’t have to be unique and they can be changed at any time, so they are not immutable.

[02:23](javascript:;)Google Cloud also assigns each project a unique project number.

[02:27](javascript:;)It’s helpful to know that these Google-generated numbers exist, but we won’t explore them much in this course.

[02:32](javascript:;)They’re mainly used internally by Google Cloud to keep track of resources.

[02:38](javascript:;)Google Cloud’s Resource Manager tool is designed to programmatically help you manage projects.

[02:43](javascript:;)It’s an API that can gather a list of all the projects associated with an account, create new projects, update existing projects, and delete projects.

[02:53](javascript:;)It can even recover projects that were previously deleted,and can be accessed through the RPC API and the REST API.

[03:02](javascript:;)The third level of the Google Cloud resource hierarchy is folders.

[03:06](javascript:;)Folders let you assign policies to resources at a level of granularity you choose.

[03:11](javascript:;)The resources in a folder inherit policies and permissions assigned to that folder.

[03:16](javascript:;)A folder can contain projects, other folders, or a combination of both.

[03:22](javascript:;)You can use folders to group projects under an organization in a hierarchy.

[03:26](javascript:;)For example, your organization might contain multiple departments, each with its own set Google Cloud resources.

[03:33](javascript:;)Folders allow you to group these resources on a per-department basis.

[03:38](javascript:;)Folders also give teams the ability to delegate administrative rights so that they can work independently.

[03:44](javascript:;)As previously mentioned, the resources in a folder inherit policies and permissions from that folder.

[03:51](javascript:;)For example, if you have two different projects that are administered by the same team, you can put policies into a common folder so they have the same permissions.

[04:01](javascript:;)Doing it the other way--putting duplicate copies of those policies on both projects–could be tedious and error-prone.

[04:08](javascript:;)if you needed to change permissions on both resources, you would now have to do that in two places instead of just one.

[04:15](javascript:;)To use folders, you must have an organization node, which is the very topmost resource in the Google Cloud hierarchy.

[04:21](javascript:;)Everything else attached to that account goes under this node, which includes folders, projects, and other resources.

[04:30](javascript:;)There are some special roles associated with this top-level organization node.

[04:34](javascript:;)For example, you can designate an organization policy administrator so that only people with privilege can change policies.

[04:43](javascript:;)You can also assign a project creator role, which is a great way to control who can create projects and, therefore, who can spend money.

[04:50](javascript:;)How a new organization node is created depends on whether your company is also a Google Workspace customer.

[04:56](javascript:;)If you have a Workspace domain, Google Cloud projects will automatically belong to your organization node.

[05:03](javascript:;)Otherwise, you can use Cloud Identity, Google’s identity, access, application, and endpoint management platform, to generate one.

[05:11](javascript:;)Once created, a new organization node will let anyone in the domain create projects and billing accounts, just as they could before.

[05:20](javascript:;)folders underneath it and put projects into it.

[05:23](javascript:;)Both folders and projects are considered to be “children” of the organization node.

Identity and Access Management (IAM)-

[00:00](javascript:;)When an organization node contains lots of folders, projects, and resources, a workforce might need to restrict who has access to what.

[00:09](javascript:;)To help with this task, administrators can use Identity and Access Management, or IAM.

[00:14](javascript:;)With IAM, administrators can apply policies that define who can do what and on which resources.

[00:21](javascript:;)The “who” part of an IAM policy can be a Google account, a Google group, a service account, or a Cloud Identity domain.

[00:29](javascript:;)A “who” is also called a “principal.”

[00:32](javascript:;)Each principle has its own identifier, usually an email address.

[00:36](javascript:;)The “can do what” part of an IAM policy is defined by a role.

[00:41](javascript:;)An IAM role is a collection of permissions.

[00:44](javascript:;)When you grant a role to a principal, you grant all the permissions that the role contains.

[00:50](javascript:;)For example, to manage virtual machine instances in a project, you must be able to create, delete, start, stop and change virtual machines.

[01:00](javascript:;)So these permissions are grouped into a role to make them easier to understand and easier to manage.

[01:07](javascript:;)When a principal is given a role on a specific element of the resource hierarchy, the

[01:11](javascript:;)resulting policy applies to both the chosen element and all the elements below it in the hierarchy.

[01:18](javascript:;)You can define deny rules that prevent certain principals from using certain permissions, regardless of the roles they're granted.

[01:25](javascript:;)This is because IAM always checks relevant deny policies before checking relevant allow policies.

[01:32](javascript:;)Deny policies, like allow policies, are inherited through the resource hierarchy.

[01:37](javascript:;)There are three kinds of roles in IAM: basic, predefined, and custom.

[01:44](javascript:;)The first role type is basic.

[01:47](javascript:;)Basic roles are quite broad in scope.

[01:50](javascript:;)When applied to a Google Cloud project, they affect all resources in that project.

[01:55](javascript:;)Basic roles include owner, editor, viewer, and billing administrator.

[02:00](javascript:;)Let’s look at these basic roles in a bit more detail.

[02:05](javascript:;)Project viewers can access resources but can’t make changes.

[02:09](javascript:;)Project editors can access and make changes to a resource.

[02:12](javascript:;)And project owners can also access and make changes to a resource.

[02:17](javascript:;)In addition, project owners can manage the associated roles and permissions and set up billing.

[02:24](javascript:;)Often companies want someone to control the billing for a project but not be able to change the resources in the project.

[02:31](javascript:;)This is possible through a billing administrator role.

[02:34](javascript:;)A word of caution: If several people are working together on a project that contains sensitive data, basic roles are probably too broad.

[02:43](javascript:;)Fortunately, IAM provides other ways to assign permissions that are more specifically tailored to meet the needs of typical job roles.

[02:51](javascript:;)This brings us to the second type of role, predefined roles.

[02:56](javascript:;)Specific Google Cloud services offer sets of predefined roles, and they even define where those roles can be applied.

[03:03](javascript:;)Let’s look at Compute Engine, for example, a Google Cloud product that offers virtual machines as a service.

[03:11](javascript:;)With Compute Engine, you can apply specific predefined roles—such as “instanceAdmin”—to Compute Engine resources in a given project, a given folder, or an entire organization.

[03:23](javascript:;)This then allows whoever has these roles to perform a specific set of predefined actions.

[03:29](javascript:;)But what if you need to assign a role that has even more specific permissions?

[03:32](javascript:;)That’s when you’d use a custom role.

[03:35](javascript:;)Many companies use a “least-privilege” model in which each person in your organization is given the minimal amount of privilege needed to do their job.

[03:44](javascript:;)So, for example, maybe you want to define an “instanceOperator” role to allow some users to stop and start Compute Engine virtual machines, but not reconfigure them.

[03:54](javascript:;)Custom roles will allow you to define those exact permissions.

[03:58](javascript:;)Before you start creating custom roles, please note two important details.

[04:02](javascript:;)First, you’ll need to manage the permissions that define the custom role you’ve created.

[04:08](javascript:;)Because of this, some organizations decide they’d rather use the predefined roles.

[04:12](javascript:;)And second, custom roles can only be applied to either the project level or organization level.

[04:18](javascript:;)They can’t be applied to the folder level.

Service Accounts-

[00:00](javascript:;)What if you want to give permissions to a Compute Engine virtual machine, rather than to a person?

[00:05](javascript:;)Well, that’s what service accounts are for.

[00:08](javascript:;)Let’s say you have an application running in a virtual machine that needs to store data in Cloud

[00:12](javascript:;)Storage, but you don’t want anyone on the internet to have access to that data–just that particular virtual machine.

[00:20](javascript:;)You can create a service account to authenticate that VM to Cloud Storage.

[00:26](javascript:;)Service accounts are named with an email address, but instead of passwords they use cryptographic keys to access resources.

[00:32](javascript:;)So, if a service account has been granted Compute Engine’s Instance Admin role, this would allow

[00:38](javascript:;)an application running in a VM with that service account to create, modify, and delete other VMs.

[00:46](javascript:;)Service accounts do need to be managed.

[00:49](javascript:;)For example, maybe Alice needs to manage which Google accounts can act as service accounts, while Bob just needs to be able to view a list of service accounts.

[00:58](javascript:;)Fortunately, in addition to being an identity, a service account is also a resource, so it can have IAM policies of its own attached to it.

[01:08](javascript:;)This means that Alice can have the editor role on a service account, and Bob can have the viewer role.

[01:13](javascript:;)This is just like granting roles for any other Google Cloud resource.

Cloud Identity-

[00:00](javascript:;)When new Google Cloud customers start using the platform, it’s common to log in to the Google Cloud

[00:04](javascript:;)Console with a Gmail account and then use Google Groups to collaborate with teammates who are in similar roles.

[00:12](javascript:;)Although this approach is easy to start with, it can present challenges later because the team’s identities are not centrally managed.

[00:19](javascript:;)This can be problematic if, for example, someone leaves the organization.

[00:24](javascript:;)With this setup, there’s no easy way to immediately remove a user’s access to the team’s cloud resources.

[00:32](javascript:;)With a tool called Cloud Identity, organizations can define policies and manage their users and groups using the Google Admin Console.

[00:41](javascript:;)Admins can log in and manage Google Cloud resources using the same usernames and passwords they already use in existing Active Directory or LDAP systems.

[00:51](javascript:;)Using Cloud Identity also means that when someone leaves an organization, an administrator can use the Google Admin Console to disable their account and remove them from groups.

[01:00](javascript:;)Cloud Identity is available in a free edition and also in a premium edition that provides capabilities to manage mobile devices.

[01:08](javascript:;)If you’re a Google Cloud customer who is also a Google Workspace customer, this functionality is already available to you in the Google Admin Console.

Interacting with Google Cloud-

[00:00](javascript:;)There are four ways to access and interact with Google Cloud.

[00:04](javascript:;)The Cloud Console, the Cloud SDK and Cloud Shell, the APIs, and the Cloud Console Mobile App.

[00:12](javascript:;)Let’s explore each of those now.

[00:16](javascript:;)First is the Google Cloud Console, which is Google Cloud’s Graphical User Interface, GUI, that helps you deploy, scale, and diagnose production issues in a simple web-based interface.

[00:29](javascript:;)With the Cloud Console, you can easily find your resources, check their health, have full management control over them, and set budgets to control how much you spend on them.

[00:40](javascript:;)The Cloud Console also provides a search facility to quickly find resources and connect to instances via SSH in the browser.

[00:49](javascript:;)Second is through the Cloud SDK and Cloud Shell.

[00:52](javascript:;)The Cloud SDK is a set of tools that you can use to manage resources and applications hosted on Google Cloud.

[00:59](javascript:;)These include the Google Cloud CLI, which provides the main command-line interface for Google Cloud products and services,

[01:05](javascript:;)gcloud storage, which lets you access Cloud Storage from the command line, and bq, a command-line tool for BigQuery.

[01:16](javascript:;)When installed, all of the tools within the Cloud SDK are located under the bin directory.

[01:23](javascript:;)Cloud Shell provides command-line access to cloud resources directly from a browser.

[01:28](javascript:;)Cloud Shell is a Debian-based virtual machine with a persistent 5 gigabyte home directory, which makes it easy to manage Google Cloud projects and resources.

[01:38](javascript:;)With Cloud Shell, the Cloud SDK gcloud command and other utilities are always installed, available, up to date, and fully authenticated.

[01:49](javascript:;)The third way to access Google Cloud is through application programming interfaces, or APIs.

[01:55](javascript:;)The services that make up Google Cloud offer APIs so that code you write can control them.

[02:01](javascript:;)The Cloud Console includes a tool called the Google APIs Explorer that shows which APIs are available, and in which versions.

[02:09](javascript:;)You can try these APIs interactively, even those that require user authentication.

[02:15](javascript:;)Suppose you’ve explored an API, and you’re ready to build an application that uses it.

[02:20](javascript:;)Do you have to start coding from scratch?

[02:22](javascript:;)No.

[02:23](javascript:;)Google provides Cloud Client libraries and Google API Client libraries in many popular languages to take

[02:28](javascript:;)a lot of the drudgery out of the task of calling Google Cloud from your code.

[02:34](javascript:;)Languages currently represented in these libraries are Java, Python, PHP, C#, Go, Node.js, Ruby, and C++.

[02:49](javascript:;)And finally, the fourth way to access and interact with Google Cloud is with the Cloud Console Mobile App, which

[02:54](javascript:;)can be used to start, stop, and use SSH to connect to Compute Engine instances and see logs from each instance.

[03:04](javascript:;)It also lets you stop and start Cloud SQL instances.

[03:08](javascript:;)Additionally, you can administer applications deployed on App Engine by viewing errors, rolling back deployments, and changing traffic splitting.

[03:18](javascript:;)The Cloud Console Mobile App provides up-to-date billing information for your projects and billing alerts for projects that are going over budget.

[03:26](javascript:;)You can set up customizable graphs showing key metrics such as CPU usage, network usage, requests per second, and server errors.

[03:37](javascript:;)The mobile app also offers alerts and incident management.

[03:42](javascript:;)You can download the Cloud Console Mobile App at cloud.google.com/console-app.

Google Cloud Fundamentals: Getting Started with Cloud Marketplace-

Lab Completed-Deploying LAMP Stack to Compute Engine

Quiz: Resource and Access in Google Cloud-

Your company has two Google Cloud projects and you want them to share policies. What is the least error-prone way to set this up?

Place both projects into a folder, and define the policies on that folder.

How does the resource hierarchy control how IAM policies are inherited?

IAM policies that are implemented by lower-level policies can override the policies defined at a higher level.

Consider a single hierarchy of Google Cloud resources. Which of these situations is possible? (Choose 3 responses.)  
There is no organization node, and there are no folders.

There is an organization node, and there are no folders.

There is an organization node, and there is at least one folder.

What is the difference between Identity and Access Management (IAM) basic roles and IAM predefined roles?

Basic roles affect all resources in a Google Cloud project. Predefined roles apply to a specific service in a project.

Which way of accessing Google Cloud lets you control services through the code you write?

APIs

Which statement best describes how Google Cloud resources are associated within the resource hierarchy?

All Google Cloud resources are associated with a project.

When would you choose to have an organization node? (Select two)

When you want to create folders

When you want to centrally apply organization-wide policies

Select the option that displays IAM roles from general to specific.

Basic roles, predefined roles, custom roles

VIRTUAL MACHINES AND NETWORKS IN THE CLOUD:

Virtual Private Cloud Networking-

[00:00](javascript:;)In this section of the course, we’re going to explore how Google Compute Engine works with a focus on virtual networking.

[00:07](javascript:;)Many users start with Google Cloud by defining their own virtual private cloud inside their first Google Cloud project or by starting with the default virtual private cloud.

[00:17](javascript:;)So, what is a virtual private cloud?

[00:21](javascript:;)A virtual private cloud, or VPC, is a secure, individual, private cloud-computing model hosted within a public cloud – like Google Cloud!

[00:32](javascript:;)On a VPC, customers can run code, store data, host websites, and do anything else they could

[00:37](javascript:;)do in an ordinary private cloud, but this private cloud is hosted remotely by a public cloud provider.

[00:46](javascript:;)This means that VPCs combine the scalability and convenience of public cloud computing with the data isolation of private cloud computing.

[00:58](javascript:;)VPC networks connect Google Cloud resources to each other and to the internet.

[01:02](javascript:;)This includes segmenting networks, using firewall rules to restrict access to instances, and creating static routes to forward traffic to specific destinations.

[01:13](javascript:;)Here's something that tends to surprise a lot of new Google Cloud users: Google VPC networks are global.

[01:21](javascript:;)They can also have subnets, which is a segmented piece of the larger network, in any Google Cloud region worldwide.

[01:29](javascript:;)Subnets can span the zones that make up a region.

[01:32](javascript:;)This architecture makes it easy to define network layouts with global scope.

[01:37](javascript:;)Resources can even be in different zones on the same subnet.

[01:42](javascript:;)The size of a subnet can be increased by expanding the range of IP addresses allocated to it, and doing so won’t affect virtual machines that are already configured.

[01:51](javascript:;)For example, let’s take a VPC with one network that currently has one subnet defined in Google Cloud’s us-east1 region.

[02:02](javascript:;)If the VPC has two Compute Engine VMs attached to it, it means they’re neighbors on the same subnet even though they’re in different zones!

Compute Engine-

[00:00](javascript:;)Earlier in the course, we explored infrastructure as a service, or IaaS.

[00:04](javascript:;)Now let’s explore Google Cloud’s IaaS solution: Compute Engine.

[00:10](javascript:;)With Compute Engine, users can create and run virtual machines on Google infrastructure.

[00:15](javascript:;)There are no upfront investments, and thousands of virtual CPUs can run on a system that’s designed to be fast and to offer consistent performance.

[00:25](javascript:;)Each virtual machine contains the power and functionality of a full-fledged operating system.

[00:31](javascript:;)This means a virtual machine can be configured much like a physical server: by specifying the amount

[00:36](javascript:;)of CPU power and memory needed, the amount and type of storage needed, and the operating system.

[00:43](javascript:;)A virtual machine instance can be created via the Google Cloud console, which is a web-based

[00:48](javascript:;)tool to manage Google Cloud projects and resources, the Google Cloud CLI, or the Compute Engine API.

[00:56](javascript:;)The instance can run Linux and Windows Server images provided by Google or any customized versions of these images.

[01:04](javascript:;)You can also build and run images of other operating systems and flexibly reconfigure virtual machines.

[01:11](javascript:;)A quick way to get started with Google Cloud is through the Cloud Marketplace, which offers solutions from both Google and third-party vendors.

[01:19](javascript:;)With these solutions, there’s no need to manually configure the software, virtual machine instances, storage, or network settings, although many of them can be modified before launch if that’s required.

[01:31](javascript:;)Most software packages in Cloud Marketplace are available at no additional charge beyond the normal usage fees for Google Cloud resources.

[01:39](javascript:;)Some Cloud Marketplace images charge usage fees, particularly those published by third parties, with commercially licensed software, but they all show estimates of their monthly charges before they’re launched.

[01:50](javascript:;)At this point, you might be wondering about Compute Engine’s pricing and billing structure.

[01:55](javascript:;)For the use of virtual machines, Compute Engine bills by the second with a one-minute minimum, and sustained-use discounts start to apply automatically to virtual machines the longer they run.

[02:07](javascript:;)So, for each VM that runs for more than 25% of a month, Compute Engine automatically applies a discount for every additional minute.

[02:16](javascript:;)Compute Engine also offers committed-use discounts.

[02:20](javascript:;)This means that for stable and predictable workloads, a specific amount of vCPUs and memory can be purchased for up to

[02:26](javascript:;)a 57% discount off of normal prices in return for committing to a usage term of one year or three years.

[02:35](javascript:;)And then there are Preemptible and Spot VMs.

[02:38](javascript:;)Let’s say you have a workload that doesn’t require a human to sit and wait for it to finish–such as a batch job analyzing a large dataset.

[02:46](javascript:;)You can save money, in some cases up to 90%, by choosing Preemptible or Spot VMs to run the job.

[02:54](javascript:;)A Preemptible or Spot VM is different from an ordinary Compute Engine VM in only

[02:57](javascript:;)one respect: Compute Engine has permission to terminate a job if its resources are needed elsewhere.

[03:06](javascript:;)Although savings are possible with preemptible or spot VMs, you'll need to ensure that your job can be stopped and restarted.

[03:12](javascript:;)Spot VMs differ from Preemptible VMs by offering more features.

[03:16](javascript:;)For example, preemptible VMs can only run for up to 24 hours at a time, but Spot VMs do not have a maximum runtime.

[03:24](javascript:;)However, the pricing is, currently the same for both.

[03:28](javascript:;)In terms of storage, Compute Engine doesn’t require a particular option or machine type to get high throughput between processing and persistent disks.

[03:36](javascript:;)That’s the default, and it comes to you at no extra cost.

[03:40](javascript:;)And finally, you’ll only pay for what you need with custom machine types.

[03:45](javascript:;)Compute Engine lets you choose the machine properties of your instances, like the number of virtual CPUs and the

[03:50](javascript:;)amount of memory, by using a set of predefined machine types or by creating your own custom machine types.

Scaling Virtual Machines-

[00:00](javascript:;)As we’ve just seen, with Compute Engine, you can choose the most appropriate machine properties for your instances, like the number

[00:06](javascript:;)of virtual CPUs and the amount of memory, by using a set of predefined machine types, or by creating custom machine types.

[00:16](javascript:;)To do this, Compute Engine has a feature called Autoscaling, where VMs can be added to or subtracted from an application based on load metrics.

[00:26](javascript:;)The other part of making that work is balancing the incoming traffic among the VMs.

[00:30](javascript:;)Google’s Virtual Private Cloud (VPC) supports several different kinds of load balancing, which we’ll explore shortly.

[00:39](javascript:;)With Compute Engine, you can in fact configure very large VMs, which are great for workloads such

[00:44](javascript:;)as in-memory databases and CPU-intensive analytics, but most Google Cloud customers start off with scaling out, not up.

[00:53](javascript:;)The maximum number of CPUs per VM is tied to its “machine family” and is also constrained by the quota available to the user, which is zone-dependent.

[01:03](javascript:;)Specifications for currently available VM machine types can be found at cloud.google.com/compute/docs/machine-types

Important VPC Capabilities-

[00:00](javascript:;)Now let’s explore some of the most important Virtual Private Cloud compatibility features.

[00:06](javascript:;)Much like physical networks, VPCs have routing tables.

[00:10](javascript:;)VPC routing tables are built-in so you don’t have to provision or manage a router.

[00:15](javascript:;)They’re used to forward traffic from one instance to another within the same network, across subnetworks, or even between Google Cloud zones, without requiring an external IP address.

[00:27](javascript:;)Another thing you don’t have to provision or manage for Google Cloud is a firewall.

[00:32](javascript:;)VPCs provide a global distributed firewall, which can be controlled to restrict access to instances through both incoming and outgoing traffic.

[00:41](javascript:;)Firewall rules can be defined through network tags on Compute Engine instances, which is really convenient.

[00:47](javascript:;)For example, you can tag all your web servers with, say, “WEB,” and write a firewall rule saying that traffic on

[00:53](javascript:;)ports 80 or 443 is allowed into all VMs with the “WEB” tag, no matter what their IP address happens to be.

[01:01](javascript:;)You’ll remember that VPCs belong to Google Cloud projects, but what if your company has several Google Cloud projects, and the VPCs need to talk to each other?

[01:12](javascript:;)With VPC Peering, a relationship between two VPCs can be established to exchange traffic.

[01:18](javascript:;)Alternatively, to use the full power of Identity Access Management (IAM) to control who and what

[01:24](javascript:;)in one project can interact with a VPC in another, you can configure a Shared VPC.

Cloud Load Balancing-

[00:00](javascript:;)Previously, we explored how virtual machines can autoscale to respond to changing loads.

[00:06](javascript:;)But how do your customers get to your application when it might be provided by four VMs one moment, and by 40 VMs at another?

[00:14](javascript:;)That’s done through Cloud Load Balancing.

[00:17](javascript:;)The job of a load balancer is to distribute user traffic across multiple instances of an application.

[00:23](javascript:;)By spreading the load, load balancing reduces the risk that applications experience performance issues.

[00:29](javascript:;)Cloud Load Balancing is a fully distributed, software-defined, managed service for all your traffic.

[00:35](javascript:;)And because the load balancers don’t run in VMs that you have to manage, you don’t have to worry about scaling or managing them.

[00:42](javascript:;)You can put Cloud Load Balancing in front of all of your traffic: HTTP or HTTPS, other TCP and SSL traffic, and UDP traffic too.

[00:52](javascript:;)Cloud Load Balancing provides cross-region load balancing, including automatic multi-region failover, which gently moves traffic in fractions if backends become unhealthy.

[01:02](javascript:;)Cloud Load Balancing reacts quickly to changes in users, traffic, network, backend health, and other related conditions.

[01:10](javascript:;)And what if you anticipate a huge spike in demand?

[01:13](javascript:;)Say, your online game is already a hit; do you need to file a support ticket to warn Google of the incoming load?

[01:20](javascript:;)No.

[01:21](javascript:;)No so-called “pre-warming” is required.

[01:24](javascript:;)VPC offers a suite of load-balancing options: If you need cross-regional load balancing for a web application, use Global HTTP(S) load balancing.

[01:33](javascript:;)For Secure Sockets Layer traffic that is not HTTP, use the Global SSL Proxy load balancer.

[01:40](javascript:;)If it’s other TCP traffic that doesn’t use SSL, use the Global TCP Proxy load balancer.

[01:46](javascript:;)Those last two proxy services only work for specific port numbers, and they only work for TCP.

[01:52](javascript:;)If you want to load balance UDP traffic, or traffic on any port number, you can still load balance across a Google Cloud region with the Regional load balancer.

[02:00](javascript:;)Finally, what all those services have in common is that they’re intended for traffic coming into the Google network from the internet.

[02:07](javascript:;)But what if you want to load balance traffic inside your project, say, between the presentation layer and the business layer of your application?

[02:15](javascript:;)For that, use the Regional internal load balancer.

[02:19](javascript:;)It accepts traffic on a Google Cloud internal IP address and load balances it across Compute Engine VMs.

Cloud DNS and Cloud CDN-

[00:00](javascript:;)One of the most famous free Google services is 8.8.8.8, which provides a public Domain Name Service to the world.

[00:07](javascript:;)DNS is what translates internet hostnames to addresses, and as you might imagine, Google has a highly developed DNS infrastructure.

[00:16](javascript:;)It makes 8.8.8.8 available so that everyone can take advantage of it.

[00:21](javascript:;)But what about the internet hostnames and addresses of applications built in Google Cloud?

[00:26](javascript:;)Google Cloud offers Cloud DNS to help the world find them.

[00:30](javascript:;)It’s a managed DNS service that runs on the same infrastructure as Google.

[00:35](javascript:;)It has low latency and high availability, and it’s a cost-effective way to make your applications and services available to your users.

[00:44](javascript:;)The DNS information you publish is served from redundant locations around the world.

[00:49](javascript:;)Cloud DNS is also programmable.

[00:51](javascript:;)You can publish and manage millions of DNS zones and records using the Cloud Console, the command-line interface, or the API.

[01:00](javascript:;)Google also has a global system of edge caches.

[01:03](javascript:;)Edge caching refers to the use of caching servers to store content closer to end users.

[01:09](javascript:;)You can use this system to accelerate content delivery in your application by using Cloud CDN - Content Delivery Network.

[01:17](javascript:;)This means your customers will experience lower network latency, the origins of your content will experience reduced load, and you can even save money.

[01:27](javascript:;)After HTTP(S) Load Balancing is set up, Cloud CDN can be enabled with a single checkbox.

[01:34](javascript:;)There are many other CDNs available out there, of course.

[01:37](javascript:;)If you are already using one, chances are, it’s a part of Google Cloud’s CDN Interconnect partner program, and you can continue to use it.

Connecting Networks to Google VPC-

[00:00](javascript:;)Many Google Cloud customers want to connect their Google Virtual Private Clouds to other networks in their system, such as on-premises networks or networks in other clouds.

[00:09](javascript:;)There are several effective ways to accomplish this.

[00:14](javascript:;)One option is to start with a Virtual Private Network connection over the internet and use the IPsec VPN protocol to create a “tunnel” connection.

[00:23](javascript:;)To make the connection dynamic, a Google Cloud feature called Cloud Router can be used.

[00:29](javascript:;)Cloud Router lets other networks and Google VPC, exchange route information over the VPN using the Border Gateway Protocol.

[00:38](javascript:;)Using this method, if you add a new subnet to your Google VPC, your on-premises network will automatically get routes to it.

[00:45](javascript:;)But using the internet to connect networks isn't always the best option for everyone, either because of security concerns or because of bandwidth reliability.

[00:54](javascript:;)So, a second option is to consider “peering” with Google using Direct Peering.

[01:01](javascript:;)Peering means putting a router in the same public data center as a Google point of presence and using it to exchange traffic between networks.

[01:10](javascript:;)Google has more than 100 points of presence around the world.

[01:13](javascript:;)Customers who aren’t already in a point of presence can work with a partner in the Carrier Peering program to get connected.

[01:22](javascript:;)Carrier peering gives you direct access from your on-premises network through a service provider's network to Google

[01:27](javascript:;)Workspace and to Google Cloud products that can be exposed through one or more public IP addresses.

[01:34](javascript:;)One downside of peering, though, is that it isn’t covered by a Google Service Level Agreement.

[01:40](javascript:;)If getting the highest uptimes for interconnection is important, using Dedicated Interconnect would be a good solution.

[01:47](javascript:;)This option allows for one or more direct, private connections to Google.

[01:51](javascript:;)If these connections have topologies that meet Google’s specifications, they can be covered by an SLA of up to 99.99%.

[01:58](javascript:;)Also, these connections can be backed up by a VPN for even greater reliability.

[02:06](javascript:;)And the final option we’ll explore is Partner Interconnect, which provides connectivity between an on-premises network and a VPC network through a supported service provider.

[02:17](javascript:;)A Partner Interconnect connection is useful if a data center is in a physical location that can't reach

[02:23](javascript:;)a Dedicated Interconnect colocation facility, or if the data needs don’t warrant an entire 10 GigaBytes per second connection.

[02:31](javascript:;)Depending on availability needs, Partner Interconnect can be configured to support mission-critical services or applications that can tolerate some downtime.

[02:40](javascript:;)As with Dedicated Interconnect, if these connections have topologies that meet Google’s specifications, they can be covered by an SLA of up to 99.99%,

[02:48](javascript:;)but note that Google isn’t responsible for any aspects of Partner Interconnect provided by the third-party service provider, nor any issues outside of Google's network.

Getting Started with VPC Networking and Google Compute Engine-Lab Completed

Quiz: Virtual Machines and Networks in the Cloud-

Select the true statement about Google’s VPC networks and subnets.

Networks are global, and subnets are regional.

A Google Cloud customer wants to load-balance traffic among the backend VMs that form part of a multi-tier application. Which load-balancing option should this customer choose?

The regional internal load balancer

Which interconnect option is a service level agreement (SLA) available for?

Dedicated Interconnect

Preemptible VMs can offer advantages over a standard Compute Engine VM. What is a reason customers choose preemptible VMs?

To reduce cost

The per-hour price of preemptible VMs incorporates a substantial discount.

Which term describes a secure, individual, private cloud-computing model hosted within a public cloud?

Virtual private cloud (VPC)

Which statement best describes how VPC routers and firewalls work?

They are managed by Google as a built-in feature.

STORAGE IN THE CLOUD:

Google Cloud Storage Options-

[00:00](javascript:;)Every application needs to store data, like media to be streamed or perhaps even sensor data from devices, and different applications and workloads require different storage database solutions.

[00:12](javascript:;)Google Cloud has storage options for structured, unstructured, transactional, and relational data.

[00:20](javascript:;)In this section of the course, we’ll explore Google Cloud’s five core storage products: Cloud Storage, Cloud SQL, Cloud Spanner, Firestore, and Cloud Bigtable.

[00:34](javascript:;)Depending on your application, you might use one or several of these services to do the job.

Cloud Storage-

[00:00](javascript:;)Let’s begin with Cloud Storage, which is a service that offers developers and IT organizations durable and highly available object storage.

[00:09](javascript:;)But what is object storage?

[00:12](javascript:;)Object storage is a computer data storage architecture that manages data as “objects” and not

[00:17](javascript:;)as a file and folder hierarchy (file storage), or as chunks of a disk (block storage).

[00:24](javascript:;)These objects are stored in a packaged format which contains the binary form of the actual data itself, as

[00:29](javascript:;)well as relevant associated meta-data (such as date created, author, resource type, and permissions), and a globally unique identifier.

[00:40](javascript:;)These unique keys are in the form of URLs, which means object storage interacts well with web technologies.

[00:47](javascript:;)Data commonly stored as objects include video, pictures, and audio recordings.

[00:53](javascript:;)Cloud Storage is Google’s object storage product.

[00:55](javascript:;)It allows customers to store any amount of data, and to retrieve it as often as needed.

[01:00](javascript:;)It’s a fully managed scalable service that has a wide variety of uses.

[01:05](javascript:;)A few examples include serving website content, storing data for archival and disaster recovery, and distributing large data objects to end users via Direct Download.

[01:17](javascript:;)Cloud Storage’s primary use is whenever binary large-object storage (also known as a “BLOB”) is needed for online content

[01:23](javascript:;)such as videos and photos, for backup and archived data and for storage of intermediate results in processing workflows.

[01:32](javascript:;)Cloud Storage files are organized into buckets.

[01:34](javascript:;)A bucket needs a globally unique name and a specific geographic location for where it should be stored, and an ideal location for a bucket is where latency is minimized.

[01:46](javascript:;)For example, if most of your users are in Europe, you probably want to pick a

[01:49](javascript:;)European location, so either a specific Google Cloud region in Europe, or else the EU multi-region.

[01:57](javascript:;)The storage objects offered by Cloud Storage are immutable, which means that you do not edit them, but instead a new version is created with every change made.

[02:07](javascript:;)Administrators have the option to either allow each new version to completely overwrite the older one, or

[02:11](javascript:;)to keep track of each change made to a particular object by enabling “versioning” within a bucket.

[02:19](javascript:;)If you choose to use versioning, Cloud Storage will keep a detailed history of modifications -- that is, overwrites or deletes -- of all objects contained in that bucket.

[02:29](javascript:;)If you don’t turn on object versioning, by default new versions will always overwrite older versions.

[02:36](javascript:;)With object versioning enabled, you can list the archived versions of an object, restore an object to an older state, or permanently delete a version of an object, as needed.

[02:47](javascript:;)In many cases, personally identifiable information may be contained in data objects, so controlling access to stored data is essential to ensuring security and privacy are maintained.

[02:59](javascript:;)Using IAM roles and, where needed, access control lists (ACLs), organizations can conform to security best practices, which require each

[03:05](javascript:;)user to have access and permissions to only the resources they need to do their jobs, and no more than that.

[03:15](javascript:;)There are a couple of options to control user access to objects and buckets.

[03:20](javascript:;)For most purposes, IAM is sufficient.

[03:23](javascript:;)Roles are inherited from project to bucket to object.

[03:27](javascript:;)If you need finer control, you can create access control lists.

[03:31](javascript:;)Each access control list consists of two pieces of information.

[03:36](javascript:;)The first is a scope, which defines who can access and perform an action.

[03:40](javascript:;)This can be a specific user or group of users.

[03:44](javascript:;)The second is a permission, which defines what actions can be performed, like read or write.

[03:51](javascript:;)Because storing and retrieving large amounts of object data can quickly become expensive, Cloud Storage also offers lifecycle management policies.

[04:00](javascript:;)For example, you could tell Cloud Storage to delete objects older than 365 days; or to delete objects created before January

[04:08](javascript:;)1, 2013; or to keep only the 3 most recent versions of each object in a bucket that has versioning enabled.

[04:16](javascript:;)Having this control ensures that you’re not paying for more than you actually need.

Cloud Storage: Storage Classes and Data Transfer-

[00:00](javascript:;)There are four primary storage classes in Cloud Storage.

[00:04](javascript:;)The first is Standard Storage.

[00:07](javascript:;)Standard Storage is considered best for frequently accessed, or “hot,” data.

[00:11](javascript:;)It’s also great for data that’s stored for only brief periods of time.

[00:16](javascript:;)The second storage class is Nearline Storage.

[00:19](javascript:;)This is best for storing infrequently accessed data, like reading or modifying data on average once a month or less.

[00:26](javascript:;)Examples might include data backups, long-tail multimedia content, or data archiving.

[00:34](javascript:;)The third storage class is Coldline Storage.

[00:37](javascript:;)This is also a low-cost option for storing infrequently accessed data.

[00:41](javascript:;)However, as compared to Nearline Storage, Coldline Storage is meant for reading or modifying data, at most, once every 90 days.

[00:51](javascript:;)The fourth storage class is Archive Storage.

[00:54](javascript:;)This is the lowest-cost option, used ideally for data archiving, online backup, and disaster recovery.

[01:01](javascript:;)It’s the best choice for data that you plan to access less than once a year,

[01:05](javascript:;)because it has higher costs for data access and operations and a 365-day minimum storage duration.

[01:13](javascript:;)Although each of these four classes has differences, it’s worth noting there are several characteristics that apply across all of these storage classes.

[01:22](javascript:;)These include: Unlimited storage with no minimum object size requirement, Worldwide accessibility and locations, Low latency and high durability, A

[01:33](javascript:;)uniform experience, which extends to security, tools, and APIs, and, Geo-redundancy if data is stored in a multi-region or dual-region.

[01:45](javascript:;)This means placing physical servers in geographically diverse data centers to protect against catastrophic events and natural disasters, and load-balancing traffic for optimal performance.

[01:57](javascript:;)Cloud Storage has no minimum fee because you pay only for what you use, and prior provisioning of capacity isn’t necessary.

[02:04](javascript:;)And from a security perspective, Cloud Storage always encrypts data on the server side, before it’s written to disk, at no additional charge.

[02:14](javascript:;)Data traveling between a customer’s device and Google is encrypted by default using HTTPS/TLS (Transport Layer Security).

[02:24](javascript:;)Regardless of which storage class you choose, there are several ways to bring data into Cloud Storage.

[02:30](javascript:;)Many customers simply carry out their own online transfer using gcloud storage, which is the Cloud Storage command from the Cloud SDK.

[02:39](javascript:;)Data can also be moved in by using a drag and drop option in the Cloud Console, if accessed through the Google Chrome web browser.

[02:47](javascript:;)But what if you have to upload terabytes or even petabytes of data?

[02:52](javascript:;)Storage Transfer Service enables you to import large amounts of online data into Cloud Storage quickly and cost-effectively.

[02:59](javascript:;)The Storage Transfer Service lets you schedule and manage batch transfers to Cloud Storage from another cloud provider, from a different Cloud Storage region, or from an HTTP(S) endpoint.

[03:09](javascript:;)And then there is the Transfer Appliance, which is a rackable, high-capacity storage server that you lease from Google Cloud.

[03:19](javascript:;)You connect it to your network, load it with data, and then ship it to an upload facility where the data is uploaded to Cloud Storage.

[03:27](javascript:;)You can transfer up to a petabyte of data on a single appliance.

[03:32](javascript:;)Cloud Storage’s tight integration with other Google Cloud products and services means that there are many additional ways to move data into the service.

[03:40](javascript:;)For example, you can import and export tables to and from both BigQuery and Cloud SQL.

[03:47](javascript:;)You can also store App Engine logs, Firestore backups, and objects used by App Engine applications, like images.

[03:55](javascript:;)Cloud Storage can also store instance startup scripts, Compute Engine images, and objects used by Compute Engine applications.

Cloud SQL-

[00:00](javascript:;)Google Cloud’s second core storage option is Cloud SQL.

[00:04](javascript:;)Cloud SQL offers fully managed relational databases, including MySQL, PostgreSQL, and SQL Server as a service.

[00:12](javascript:;)It’s designed to hand off mundane, but necessary and often time-consuming, tasks to Google—like applying

[00:19](javascript:;)patches and updates managing backups, and configuring replications—so your focus can be on building great applications.

[00:27](javascript:;)Cloud SQL: Doesn't require any software installation or maintenance.

[00:32](javascript:;)Can scale up to 64 processor cores, 400+ GB of RAM, and 30 TB of storage.

[00:39](javascript:;)Supports automatic replication scenarios, such as from a Cloud SQL primary instance, an external primary instance, and external MySQL instances.

[00:52](javascript:;)Supports managed backups, so backed-up data is securely stored and accessible if a restore is required.

[00:58](javascript:;)The cost of an instance covers seven backups.

[01:01](javascript:;)Encrypts customer data when on Google’s internal networks and when stored in database tables, temporary files, and backups.

[01:09](javascript:;)Includes a network firewall, which controls network access to each database instance.

[01:15](javascript:;)A benefit of Cloud SQL instances is that they are accessible by other Google Cloud services, and even external services.

[01:24](javascript:;)Cloud SQL can be used with App Engine using standard drivers like Connector/J for Java or MySQLdb for Python.

[01:33](javascript:;)Compute Engine instances can be authorized to access Cloud SQL instances and configure the Cloud SQL instance to be in the same zone as your virtual machine.

[01:43](javascript:;)Cloud SQL also supports other applications and tools that you might use, like SQL Workbench, Toad, and other external applications using standard MySQL drivers.

Cloud Spanner-

[00:00](javascript:;)The third core storage option offered by Google Cloud is Cloud Spanner.

[00:04](javascript:;)Cloud Spanner is a fully managed relational database service that scales horizontally, is strongly consistent, and speaks SQL.

[00:12](javascript:;)Battle tested by Google’s own mission-critical applications and services, Spanner is the service that powers Google’s $80 billion business.

[00:20](javascript:;)Cloud Spanner is especially suited for applications that require: A SQL relational database management system with joins and

[00:28](javascript:;)secondary indexes Built-in high availability Strong global consistency And high numbers of input and output operations per second.

[00:38](javascript:;)We’re talking tens of thousands of reads and writes per second or more.

FireStore-

[00:00](javascript:;)Google Cloud’s fourth core storage option is Firestore.

[00:04](javascript:;)Firestore is a flexible, horizontally scalable, NoSQL cloud database for mobile, web, and server development.

[00:13](javascript:;)With Firestore, data is stored in documents and then organized into collections.

[00:18](javascript:;)Documents can contain complex nested objects in addition to subcollections.

[00:22](javascript:;)Firestore’s NoSQL queries can then be used to retrieve individual, specific documents or to retrieve all the documents in a collection that match your query parameters.

[00:34](javascript:;)Queries can include multiple, chained filters and combine filtering and sorting options.

[00:39](javascript:;)They're also indexed by default, so query performance is proportional to the size of the result set, not the dataset.

[00:47](javascript:;)Firestore uses data synchronization to update data on any connected device.

[00:52](javascript:;)However, it's also designed to make simple, one-time fetch queries efficiently.

[00:58](javascript:;)It caches data that an app is actively using, so the app can write, read, listen to, and query data even if the device is offline.

[01:07](javascript:;)When the device comes back online, Firestore synchronizes any local changes back to Firestore.

[01:13](javascript:;)Firestore leverages Google Cloud’s powerful infrastructure: automatic multi-region data replication, strong consistency guarantees, atomic batch operations, and real transaction support.

[01:27](javascript:;)From a pricing perspective, you’re charged for each document read, write, and delete that you perform with Firestore.

[01:34](javascript:;)Queries are also charged at the rate of one “document read” per query, whether the query returns data or not.

[01:40](javascript:;)You’re also charged for the amount of storage your data consumes and for certain kinds of network bandwidth used to access your data.

[01:47](javascript:;)Ingress is currently free, and in many cases so is egress.

[01:52](javascript:;)Consult the pricing page for Firestore for details, or you can use Google’s Billing Calculator to estimate prices for your particular use case.

[02:01](javascript:;)In addition to the 10GiB of free network egress per month between US regions, Firestore has a free

[02:07](javascript:;)quota per day of: 50,000 document reads, 20,000 document writes, 20,000 document deletes, and 1 GB of stored data.

[02:19](javascript:;)Charges only begin once the free daily quota has been exceeded.

[02:23](javascript:;)This allows you to get started developing with Firestore for very little, or even for free.

Cloud Bigtable-

[00:00](javascript:;)The last of Google Cloud’s core storage options we’re going to explore is Cloud Bigtable.

[00:06](javascript:;)Cloud Bigtable is Google's NoSQL big data database service.

[00:11](javascript:;)It's the same database that powers many core Google services, including Search, Analytics, Maps, and Gmail.

[00:18](javascript:;)Bigtable is designed to handle massive workloads at consistent low latency and high throughput, so it's a

[00:24](javascript:;)great choice for both operational and analytical applications, including Internet of Things, user analytics, and financial data analysis.

[00:33](javascript:;)When deciding which storage option is best, customers often choose Bigtable if: They’re working with more than 1TB of semi-structured or structured data.

[00:43](javascript:;)Data is fast with high throughput, or it’s rapidly changing.

[00:46](javascript:;)They’re working with NoSQL data.

[00:49](javascript:;)This usually means transactions where strong relational semantics are not required.

[00:55](javascript:;)Data is a time-series or has natural semantic ordering.

[00:58](javascript:;)They’re working with big data, running asynchronous batch or synchronous real-time processing on the data.

[01:04](javascript:;)Or they’re running machine learning algorithms on the data.

[01:09](javascript:;)Cloud Bigtable can interact with other Google Cloud services and third-party clients.

[01:15](javascript:;)Using APIs, data can be read from and written to Cloud Bigtable through a data service

[01:19](javascript:;)layer like Managed VMs, the HBase REST Server, or a Java Server using the HBase client.

[01:28](javascript:;)Typically this is used to serve data to applications, dashboards, and data services.

[01:35](javascript:;)Data can also be streamed in through a variety of popular stream processing frameworks like Dataflow Streaming, Spark Streaming, and Storm.

[01:44](javascript:;)And if streaming is not an option, data can also be read from and written to Cloud Bigtable through batch processes like Hadoop MapReduce, Dataflow, or Spark.

[01:54](javascript:;)Often, summarized or newly calculated data is written back to Cloud Bigtable or to a downstream database.

Comparing Storage Options-

[00:00](javascript:;)Now that we’ve covered Google Cloud’s core storage options, let’s do a comparison to help highlight the most suitable service for a specific application or workflow.

[00:10](javascript:;)Consider using Cloud Storage if you need to store immutable blobs larger than 10 megabytes, such as large images or movies.

[00:18](javascript:;)This storage service provides petabytes of capacity with a maximum unit size of 5 terabytes per object.

[00:26](javascript:;)Consider using Cloud SQL or Cloud Spanner if you need full SQL support for an online transaction processing system.

[00:33](javascript:;)Cloud SQL provides up to 64 terabytes, depending on machine type, and Cloud Spanner provides petabytes.

[00:40](javascript:;)Cloud SQL is best for web frameworks and existing applications, like storing user credentials and customer orders.

[00:48](javascript:;)If Cloud SQL doesn’t fit your requirements because you need horizontal scalability, not just through read replicas, consider using Cloud Spanner.

[00:57](javascript:;)Consider Firestore if you need massive scaling and predictability together with real time query results and offline query support.

[01:05](javascript:;)This storage service provides terabytes of capacity with a maximum unit size of 1 megabyte per entity.

[01:12](javascript:;)Firestore is best for storing, syncing, and querying data for mobile and web apps.

[01:18](javascript:;)Finally, consider using Cloud Bigtable if you need to store a large number of structured objects.

[01:25](javascript:;)Cloud Bigtable doesn’t support SQL queries, nor does it support multi-row transactions.

[01:31](javascript:;)This storage service provides petabytes of capacity with a maximum unit size of 10 megabytes per cell and 100 megabytes per row.

[01:39](javascript:;)Bigtable is best for analytical data with heavy read and write events, like AdTech, financial, or IoT data.

[01:48](javascript:;)Depending on your application, it’s possible that you might use one, or several, of these services to do the job.

[01:54](javascript:;)You may have noticed that BigQuery hasn’t been mentioned in this section of the course.

[01:58](javascript:;)This is because it sits on the edge between data storage and data processing, and is covered in more depth in other courses.

[02:06](javascript:;)The usual reason to store data in BigQuery is so you can use its big data analysis and interactive querying capabilities, but it’s not purely a data storage product.

Google Cloud Fundamentals: Getting Started with Cloud Storage and Cloud SQL-

Lab completed

Quiz: Storage in the Cloud-

How are Firestore and Bigtable alike? (Select two answers.)

They are both NoSQL databases.

They are both highly scalable.

Which statement describes the correct Cloud Storage use case?

Cloud Storage provides durable and highly available object storage.

Your application needs to store data with strong transactional consistency, and you want seamless scaling up. Which storage option is the best choice for your application?

Cloud Spanner

Why would a customer consider the Coldline Storage class?

To save money on storing infrequently accessed data

You manufacture devices with sensors and need to stream huge amounts of data from these devices to a storage option in the cloud. Which storage option is the best choice for your application?

Bigtable

Which SQL database service can scale to petabyte database sizes?

Cloud Spanner

CONTAINERS IN THE CLOUD:

Introduction to Containers-

[00:00](javascript:;)In this section of the course we’ll explore containers and help you understand how they are used.

[00:06](javascript:;)Infrastructure as a service, or IaaS, allows you to share compute resources with other developers by using virtual machines to virtualize the hardware.

[00:16](javascript:;)This lets each developer deploy their own operating system (OS), access the hardware, and build their applications in a self-contained environment with access to RAM, file systems, networking interfaces, etc.

[00:30](javascript:;)This is where containers come in.

[00:33](javascript:;)The idea of a container is to give the independent scalability of workloads in PaaS and an abstraction layer of the OS and hardware in IaaS.

[00:43](javascript:;)A configurable system lets you install your favorite runtime, web server, database, or middleware, configure the

[00:49](javascript:;)underlying system resources, such as disk space, disk I/O, or networking, and build as you like.

[00:57](javascript:;)But flexibility comes with a cost.

[00:59](javascript:;)The smallest unit of compute is an app with its VM.

[01:03](javascript:;)The guest OS might be large, even gigabytes in size, and take minutes to boot.

[01:08](javascript:;)As demand for your application increases, you have to copy an entire VM and boot the guest OS for each instance of your app, which can be slow and costly.

[01:18](javascript:;)Now, with App Engine, you get access to programming services, so you only need to write your code in self-contained workloads that use these services and include any dependent libraries.

[01:31](javascript:;)This means that as demand for your app increases, the platform scales your app seamlessly and independently by workload and infrastructure.

[01:39](javascript:;)This scales rapidly, but there’s no option to fine-tune the underlying architecture to save cost.

[01:46](javascript:;)A container is an invisible box around your code and its dependencies with limited access to its own partition of the file system and hardware.

[01:55](javascript:;)It only requires a few system calls to create and it starts as quickly as a process.

[02:00](javascript:;)All that’s needed on each host is an OS kernel that supports containers and a container runtime.

[02:06](javascript:;)In essence, the OS is being virtualized.

[02:09](javascript:;)It scales like PaaS but gives you nearly the same flexibility as IaaS.

[02:15](javascript:;)This makes code ultra portable, and the OS and hardware can be treated as a black box.

[02:20](javascript:;)So you can go from development, to staging, to production, or from your laptop to the cloud, without changing or rebuilding anything.

[02:30](javascript:;)As an example, let’s say you want to scale a web server.

[02:34](javascript:;)With a container, you can do this in seconds and deploy dozens or hundreds of them, depending on the size or your workload, on a single host.

[02:42](javascript:;)That's just a simple example of scaling one container running the whole application on a single host.

[02:47](javascript:;)However, you'll probably want to build your applications using lots of containers, each performing their own function like microservices.

[02:55](javascript:;)If you build them this way and connect them with network connections, you can make them modular, deploy easily, and scale independently across a group of hosts.

[03:05](javascript:;)The hosts can scale up and down and start and stop containers as demand for your app changes or as hosts fail.

Kubernetes-

[00:00](javascript:;)A product that helps manage and scale containerized applications is Kubernetes.

[00:04](javascript:;)So to save time and effort when scaling applications and workloads, Kubernetes can be bootstrapped using Google Kubernetes Engine or GKE.

[00:14](javascript:;)So, what is Kubernetes?

[00:17](javascript:;)Kubernetes is an open-source platform for managing containerized workloads and services.

[00:23](javascript:;)It makes it easy to orchestrate many containers on many hosts, scale them as microservices, and easily deploy rollouts and rollbacks.

[00:31](javascript:;)At the highest level, Kubernetes is a set of APIs that you can use to deploy containers on a set of nodes called a cluster.

[00:40](javascript:;)The system is divided into a set of primary components that run as the control plane and a set of nodes that run containers.

[00:48](javascript:;)In Kubernetes, a node represents a computing instance, like a machine.

[00:52](javascript:;)Note that this is different to a node on Google Cloud which is a virtual machine running in Compute Engine.

[00:59](javascript:;)You can describe a set of applications and how they should interact with each other, and Kubernetes determines how to make that happen.

[01:07](javascript:;)Deploying containers on nodes by using a wrapper around one or more containers is what defines a Pod.

[01:14](javascript:;)A Pod is the smallest unit in Kubernetes that you create or deploy.

[01:18](javascript:;)It represents a running process on your cluster as either a component of your application or an entire app.

[01:25](javascript:;)Generally, you only have one container per Pod, but if you have multiple containers with a hard

[01:30](javascript:;)dependency, you can package them into a single Pod and share networking and storage resources between them.

[01:37](javascript:;)The Pod provides a unique network IP and set of ports for your containers and configurable options that govern how your containers should run.

[01:45](javascript:;)One way to run a container in a Pod in Kubernetes is to use the kubectl run command, which starts a Deployment with a container running inside a Pod.

[01:55](javascript:;)A Deployment represents a group of replicas of the same Pod and keeps your Pods running even when the nodes they run on fail.

[02:03](javascript:;)A Deployment could represent a component of an application or even an entire app.

[02:09](javascript:;)To see a list of the running Pods in your project, run the command: $ kubectl get pods Kubernetes creates a Service with a fixed IP address for your

[02:17](javascript:;)Pods, and a controller says "I need to attach an external load balancer with a public IP address to that Service so others outside the cluster can access it."

[02:26](javascript:;)In GKE, the load balancer is created as a network load balancer.

[02:32](javascript:;)Any client that reaches that IP address will be routed to a Pod behind the Service.

[02:39](javascript:;)A Service is an abstraction which defines a logical set of Pods and a policy by which to access them.

[02:45](javascript:;)As Deployments create and destroy Pods, Pods will be assigned their own IP addresses, but those addresses don't remain stable over time.

[02:55](javascript:;)A Service group is a set of Pods and provides a stable endpoint (or fixed IP address) for them.

[03:01](javascript:;)For example, if you create two sets of Pods called frontend and backend and put them

[03:05](javascript:;)behind their own Services, the backend Pods might change, but frontend Pods are not aware of this.

[03:13](javascript:;)They simply refer to the backend Service.

[03:16](javascript:;)To scale a Deployment, run the kubectl scale command.

[03:20](javascript:;)In this example, three Pods are created in your Deployment, and they're placed behind the Service and share one fixed IP address.

[03:29](javascript:;)You could also use autoscaling with other kinds of parameters; for example, you can specify that the number of Pods should increase when CPU utilization reaches a certain limit.

[03:41](javascript:;)So far, we’ve seen how to run imperative commands like expose and scale.

[03:46](javascript:;)This works well to learn and test Kubernetes step-by-step.

[03:48](javascript:;)But the real strength of Kubernetes comes when you work in a declarative way.

[03:54](javascript:;)Instead of issuing commands, you provide a configuration file that tells Kubernetes what you want your desired state to look like, and Kubernetes determines how to do it.

[04:04](javascript:;)You accomplish this by using a Deployment config file.

[04:08](javascript:;)To get this file, you can run a kubectl get deployments command, and you'll get a Deployment configuration file that looks like this.

[04:16](javascript:;)You can check your Deployment to make sure the proper number of replicas is running by using either kubectl get deployments or kubectl describe deployments.

[04:26](javascript:;)To run five replicas instead of three, all you do is update the Deployment config file and run the kubectl apply command to use the updated config file.

[04:36](javascript:;)You can still reach your endpoint as before by using kubectl get services to get the external IP of the Service and reach the public IP address from a client.

[04:46](javascript:;)The last question is, what happens when you want to update a new version of your app?

[04:50](javascript:;)Well, you want to update your container to get new code in front of users, but rolling out all those changes at one time would be risky.

[04:57](javascript:;)So in this case, you would use kubectl rollout or change your deployment configuration file and then apply the change using kubectl apply.

[05:07](javascript:;)New Pods will then be created according to your new update strategy.

[05:10](javascript:;)Here’s an example configuration that will create new version Pods individually and wait for a new Pod to be available before destroying one of the old Pods.

Google Kubernetes Engine-

[00:00](javascript:;)So now that we have a basic understanding of containers and Kubernetes, let’s talk about Google Kubernetes Engine, or GKE.

[00:07](javascript:;)GKE is a Google-hosted managed Kubernetes service in the cloud.

[00:12](javascript:;)The GKE environment consists of multiple machines, specifically Compute Engine instances, grouped together to form a cluster.

[00:21](javascript:;)You can create a Kubernetes cluster with Kubernetes Engine by using the Google Cloud Console or the g-cloud command that's provided by the Cloud software development kit.

[00:30](javascript:;)GKE clusters can be customized, and they support different machine types, number of nodes, and network settings.

[00:38](javascript:;)Kubernetes provides the mechanisms through which you interact with your cluster.

[00:43](javascript:;)Kubernetes commands and resources are used to deploy and manage applications, perform administration tasks, set policies, and monitor the health of deployed workloads.

[00:54](javascript:;)Running a GKE cluster comes with the benefit of advanced cluster management features that Google Cloud provides.

[01:00](javascript:;)These include: Google Cloud's load-balancing for Compute Engine instances Node pools to designate subsets of nodes within a cluster for

[01:09](javascript:;)additional flexibility Automatic scaling of your cluster's node instance count Automatic upgrades for your cluster's node software Node auto-repair to

[01:18](javascript:;)maintain node health and availability Logging and monitoring with Google Cloud's operations suite for visibility into your cluster Running your

[01:28](javascript:;)application in GKE clusters is also a good foundation to have if you’ll need to bridge your on-prem and cloud resources.

[01:35](javascript:;)To start up Kubernetes on a cluster in GKE, all you do is run this command: $> gcloud container clusters create k1

Hybrid and Multi-Cloud-

[00:00](javascript:;)Now that you’ve seen how containers work, we’re going to take that information a step

[00:03](javascript:;)further and explore how they can be used in a modern hybrid cloud and multi-cloud architecture.

[00:09](javascript:;)Let’s begin by looking at a typical on-premises distributed systems architecture, which is how businesses traditionally met their enterprise computing needs before cloud computing.

[00:21](javascript:;)Most enterprise-scale applications are designed as distributed systems, spreading the computing workload required to provide services over two or more networked servers.

[00:32](javascript:;)In recent years, containers have become a popular way to break these workloads down into “microservices” so they can be more easily maintained and expanded.

[00:40](javascript:;)Traditionally, these enterprise systems–and their workloads, containerized or not–have been located on-premises, which means they’re housed on

[00:47](javascript:;)a set of high-capacity servers running somewhere within the company’s network or within a company-owned data center.

[00:55](javascript:;)When an application’s computing needs begin to outstrip its available computing resources, a company using on-premises systems would need to requisition more (or more powerful) servers, install them on the

[01:06](javascript:;)company network (after any necessary network changes or expansion), configure the new servers, and finally

[01:13](javascript:;)load the application and its dependencies onto the new servers, before resource bottlenecks could be resolved.

[01:20](javascript:;)The time required to complete an on-premises upgrade of this kind could be anywhere from several months to one or more years.

[01:27](javascript:;)It might also be quite costly, especially when you consider the useful lifespan of the average server is only 3-5 years.

[01:35](javascript:;)But what if you need more computing power now, not months from now?

[01:39](javascript:;)What if your company wants to relocate some workloads away from on-premises to the cloud to take advantage of

[01:44](javascript:;)lower costs and higher availability, but is unwilling (or unable) to move the entire enterprise application from the on-premises network?

[01:53](javascript:;)And what if you want to use specialized products and services that are only available in the cloud?

[01:59](javascript:;)This is where a modern hybrid or multi-cloud architecture can help.

[02:02](javascript:;)It allows you to keep parts of your systems infrastructure on-premises while moving other parts to the cloud, thus creating an environment that is uniquely suited to your company’s needs.

[02:13](javascript:;)Move only specific workloads to the cloud at your own pace, because a full-scale migration is not required for it to work.

[02:20](javascript:;)Take advantage of the flexibility, scalability, and lower computing costs offered by cloud services for running the workloads you decide to migrate.

[02:28](javascript:;)And add specialized services, such as machine learning, content caching, data analysis, long-term storage, and IoT, to your computing resources toolkit.

[02:39](javascript:;)In the next video you’ll learn about Google Cloud’s answer to modern hybrid and multi-cloud distributed systems and services management.

Anthos-

[00:00](javascript:;)You might have heard a lot recently concerning the adoption of “hybrid” architecture for powering distributed systems and services.

[00:07](javascript:;)You might have even heard about Google’s answer to modern hybrid and multi-cloud distributed systems and services management, called Anthos.

[00:15](javascript:;)But what exactly is Anthos?

[00:18](javascript:;)Anthos is a hybrid and multi-cloud solution powered by the latest innovations in distributed systems and service management software from Google.

[00:27](javascript:;)The Anthos framework rests on Kubernetes and GKE On-Prem.

[00:31](javascript:;)This provides the foundation for an architecture that’s fully integrated and has centralized management through

[00:36](javascript:;)a central control plane that supports policy-based application lifecycle delivery across hybrid and multiple cloud environments.

[00:45](javascript:;)Anthos also provides a rich set of tools for monitoring and maintaining the consistency of

[00:49](javascript:;)your applications across all of your network, whether on-premises, in the cloud, or in multiple clouds.

[00:56](javascript:;)Let’s take a deeper look at this framework as we build a modern hybrid infrastructure stack, step by step, with Anthos.

[01:04](javascript:;)First, let’s look at Google Kubernetes Engine on the cloud side of our hybrid network.

[01:11](javascript:;)Google Kubernetes Engine: Is a managed, production-ready environment for deploying containerized applications.

[01:18](javascript:;)Operates seamlessly with high availability and an SLA.

[01:22](javascript:;)Runs Certified Kubernetes, thus ensuring portability across clouds and on-premises.

[01:29](javascript:;)Includes auto node repair, auto upgrade, and autoscaling.

[01:35](javascript:;)And uses regional clusters for high availability with multiple control planes and node storage replication across multiple zones.

[01:44](javascript:;)Its counterpart on the on-premises side of our hybrid network is GKE On-Prem.

[01:48](javascript:;)GKE On-Prem: Is a turnkey, production-grade, conformant version of Kubernetes with a best-practice configuration pre-loaded.

[01:59](javascript:;)Provides an easy upgrade path to the latest Kubernetes releases that have been validated and tested by Google.

[02:05](javascript:;)Provides access to container services on Google Cloud such as Cloud Build, Container Registry, and Cloud Audit Logs.

[02:12](javascript:;)Integrates with Istio, Knative, and Cloud Marketplace solutions.

[02:17](javascript:;)And ensures a consistent Kubernetes version and experience across cloud and on-premises environments.

[02:24](javascript:;)Both Google Kubernetes Engine and GKE On-Prem integrate with Marketplace so that all of the clusters in

[02:29](javascript:;)your network, whether on-premises or in the cloud, have access to the same repository of containerized applications.

[02:37](javascript:;)This allows you to use the same configurations on both sides of the network, which reduces time spent maintaining conformity between your clusters.

[02:45](javascript:;)You also spend less time developing applications because of a write-once/replicate-anywhere approach.

[02:52](javascript:;)Enterprise applications might use hundreds of microservices to handle computing workloads.

[02:59](javascript:;)Keeping track of all these services and monitoring their health can quickly become a challenge.

[03:05](javascript:;)Anthos Service Mesh and Istio Open Source Service Mesh take all of the guesswork out of managing and securing your microservices.

[03:13](javascript:;)These service mesh layers communicate across the hybrid network using Cloud Interconnect to sync and pass their data.

[03:21](javascript:;)Cloud Logging and Cloud Monitoring are the built-in logging and monitoring solutions for Google Cloud.

[03:26](javascript:;)Google Cloud’s operations suite offers a fully managed logging, metrics collection, monitoring, dashboarding, and alerting solution that watches all sides of your hybrid or multi-cloud network.

[03:38](javascript:;)It’s the ideal solution for customers wanting a single, easy to configure, powerful cloud-based observability solution

[03:44](javascript:;)that also gives you a ‘single pane of glass’ dashboard to monitor all of your environments.

[03:51](javascript:;)Finally, Anthos Configuration Management provides a single authoritative source of truth for your clusters’ configuration.

[04:00](javascript:;)This is kept in the Policy Repository, which is actually a git repository.

[04:05](javascript:;)The repository can be located on-premises or hosted in the cloud.

[04:10](javascript:;)The Anthos Configuration Management agents use the Policy Repository to enforce configurations locally in each environment, thus managing the complexity of owning clusters across environments.

[04:22](javascript:;)Anthos Configuration Management also gives administrators and developers the ability to deploy code changes with a single repository commit and the

[04:29](javascript:;)option to implement configuration inheritance by using Namespaces, which is a way to prevent naming and permissions collisions within your application.

[04:40](javascript:;)You can learn more about Anthos by heading to cloud.google.com/anthos.

Google Cloud Fundamentals: Getting Started with GKE- Lab Completed

Quiz: Containers in the Cloud-

How do containers access an operating system?

Containers use a shared base operating system stored in a shared kernel layer.

Select two reasons for using containers to deploy applications. (Choose 2 responses.)

It creates consistency across development, testing, and production environments.  
Migrating workloads is simpler.

Anthos provides a rich set of tools for monitoring and maintaining the consistency of your applications across which of the following locations?

Applications hosted on-premises, in the cloud, or in multiple clouds.

Where do the resources used to build Google Kubernetes Engine clusters come from?

Compute Engine

What is a Kubernetes cluster?

A group of machines where Kubernetes can schedule workloads.

How do you keep your Kubernetes version updated in Google Kubernetes Engine?

The Google Kubernetes Engine team periodically performs automatic upgrades of your cluster to newer stable versions.

What is a Kubernetes pod?

A group of containers

APPLICATIONS IN THE CLOUD:

App Engine-

[00:00](javascript:;)So far in this course, we’ve provided an introduction to Google Cloud and explored the options and benefits related to using virtual machines networks, storage, and containers in the Cloud.

[00:12](javascript:;)In this section of the course, we’ll turn our attention to developing applications in the Cloud.

[00:17](javascript:;)We’ll begin with App Engine, which is a fully managed, serverless platform for developing and hosting web applications at scale.

[00:25](javascript:;)So, how does it work?

[00:28](javascript:;)With App Engine, you can choose from popular coding languages, libraries, and frameworks to develop apps

[00:33](javascript:;)with tools you’re familiar with, and then automatically provision servers and scale app instances based on demand.

[00:41](javascript:;)This means you can upload your code and Google will manage your app's availability.

[00:45](javascript:;)Coding options include Eclipse, IntelliJ, Maven, Git, Jenkins, and PyCharm.

[00:55](javascript:;)With App Engine, there are no servers to provision or maintain.

[00:58](javascript:;)App Engine provides built-in services and APIs, like NoSQL datastores Memcache load balancing, health checks, application logging, and a user authentication API that’s common to most applications.

[01:14](javascript:;)App Engine also offers software development kits, SDKs, to help you develop, deploy, and manage your apps on your local machine.

[01:23](javascript:;)Each SDK includes: All of the APIs and libraries available to App Engine, A simulated, secure sandbox environment that emulates all of

[01:31](javascript:;)the App Engine services on your local computer, And deployment tools to upload your application to the cloud and manage different versions.

[01:41](javascript:;)The SDK manages your application locally, and the Google Cloud Console manages your application in production.

[01:49](javascript:;)You can use the Cloud Console’s web-based interface to create new applications, configure domain names,

[01:53](javascript:;)change which version of your application is live, examine access and error logs, and much more.

[02:02](javascript:;)From a security perspective, the Security Command Center–Google Cloud’s security and risk management platform–keeps web applications safe.

[02:11](javascript:;)Through the Cloud Console, you can use the Security Command Center to automatically scan and detect common web application vulnerabilities.

App Engine Environments-

[00:00](javascript:;)There are two types of App Engine environments: standard and flexible.

[00:05](javascript:;)The App Engine standard environment is based on container instances running on Google's infrastructure.

[00:11](javascript:;)Containers are preconfigured with a runtime from a standardized list of supported languages and versions, which includes libraries that support App Engine standard APIs.

[00:20](javascript:;)For many applications, the standard environment runtimes and libraries may be all you need.

[00:27](javascript:;)Standard environment features include: Persistent storage with queries, sorting, and transactions Automatic scaling and load balancing Asynchronous task queues for performing

[00:37](javascript:;)work outside the scope of a request Scheduled tasks for triggering events at specified times or regular intervals And integration with

[00:47](javascript:;)other Google Cloud services and APIs There are a couple of requirements for using the standard environment: You must use specified

[00:56](javascript:;)versions of Java, Python, PHP, Go, Node.js, and Ruby, and Your application must conform to sandbox constraints that are dependent on runtime.

[01:10](javascript:;)Applications run in a secure, sandboxed environment.

[01:13](javascript:;)This allows the App Engine standard environment to distribute requests across multiple servers and scale servers to meet traffic demands.

[01:22](javascript:;)This means that your application runs within its own secure, reliable environment that is independent of the hardware operating system, or physical location of the server.

[01:32](javascript:;)A standard environment workflow typically follows these three steps: First, a web application is developed and tested locally.

[01:41](javascript:;)Second, the SDK is used to deploy the application to App Engine.

[01:46](javascript:;)And third, App Engine scales and services the application.

[01:51](javascript:;)App Engine also offers a flexible environment.

[01:55](javascript:;)If the standard environment’s sandbox model is too restrictive for you, the flexible environment can let you specify the type of container your web application will run in.

[02:05](javascript:;)This option lets an application run inside Docker containers on Google Cloud’s Compute Engine virtual machines.

[02:10](javascript:;)In this case, App Engine manages Compute Engine machines for you.

[02:16](javascript:;)This means that: Instances are health-checked, healed as necessary, and co-located with other module instances within the project.

[02:24](javascript:;)Critical, backward-compatible updates are automatically applied to the underlying operating system.

[02:29](javascript:;)VM instances are automatically located by geographical region according to the settings in your project.

[02:36](javascript:;)Google's management services ensure that all of a project's VM instances are co-located for optimal performance.

[02:44](javascript:;)And VM instances are restarted on a weekly basis.

[02:47](javascript:;)During restarts, Google's management services will apply any necessary operating system and security updates.

[02:53](javascript:;)The flexible environment supports microservices, authorization, SQL and NoSQL databases, traffic splitting, logging, search, versioning, security scanning, Memcache, and content delivery networks.

[03:10](javascript:;)App Engine Flexible allows users to also benefit from custom configurations and libraries while still keeping their main focus on what they do best – writing code.

[03:21](javascript:;)In addition, the App Engine flexible environment allows you to customize the runtime and the operating system of your virtual machine by using Dockerfiles.

[03:31](javascript:;)As in App Engine Standard, supported runtimes include Python, Java, Go, Node.js, PHP, and Ruby.

[03:40](javascript:;)However, in App Engine Flexible, developers can also use different versions of these runtimes or provide their

[03:46](javascript:;)own custom runtime by supplying a custom Docker image or using a Dockerfile from the open source community.

[03:53](javascript:;)So, how do these two environments compare to each other?

[03:57](javascript:;)Let’s start with the standard environment, which is fast.

[04:01](javascript:;)It starts up instances of your application in seconds, but you have less access to the infrastructure in which your application runs.

[04:07](javascript:;)With the standard environment, you can’t use SSH to connect to the virtual machines on which your application runs, and you can’t write to a local disk.

[04:17](javascript:;)The standard environment does support third-party binaries for certain languages, and you can use App Engine to make calls to the network.

[04:23](javascript:;)Finally, in terms of pricing, after a free tier usage, you pay per instance class with automatic shutdown.

[04:32](javascript:;)The flexible environment takes minutes to start up, instead of seconds.

[04:35](javascript:;)But it lets you use SSH to connect to the virtual machines on which your application runs, it lets you use local disk

[04:41](javascript:;)for scratch space, it lets you install third-party software, and it lets your application make calls to the network without going through App Engine.

[04:50](javascript:;)In terms of pricing, with the flexible environment, you pay for resource allocation per hour with no automatic shutdown.

[04:58](javascript:;)Because App Engine uses Docker containers, you may be wondering how App Engine compares to Google Kubernetes Engine.

[05:04](javascript:;)App Engine's standard environment is for people who want the service to take maximum control of their web and mobile application’s deployment and scaling.

[05:06](javascript:;)App Engine's standard environment is for people who want the service to take maximum control of their web and mobile application’s deployment and scaling.

[05:14](javascript:;)Google Kubernetes Engine, however, gives the application owner the full flexibility of Kubernetes.

[05:20](javascript:;)App Engine's flexible environment is somewhere between the two.

Google Cloud API Management Tools-

[00:00](javascript:;)Now that you’ve had a thorough overview of App Engine, let’s transition to Cloud Endpoints and Apigee API Management.

[00:06](javascript:;)These are management tools for Google Cloud’s application programming interface, or API.

[00:12](javascript:;)So what exactly is an API?

[00:13](javascript:;)A software service’s implementation can be complex and changeable.

[00:19](javascript:;)If other software services had to be explicitly coded in detail in order to use that service, the result would be brittle and error-prone.

[00:27](javascript:;)So instead, application developers structure the software they write so that it presents a clean, well-defined interface that hides unnecessary detail, and then they document that interface.

[00:38](javascript:;)That’s an application programming interface.

[00:40](javascript:;)The underlying implementation can change, as long as the interface doesn’t, and other pieces of software that use the API don’t have to know or care.

[00:51](javascript:;)Sometimes you do have to change an API, perhaps to add or deprecate a feature.

[00:56](javascript:;)To cleanly make this kind of change to the API, developers create versions.

[01:01](javascript:;)For example, version 2 of an API might contain calls that version 1 does not.

[01:06](javascript:;)This means that programs that consume the API can specify the API version they want to use in their calls.

[01:13](javascript:;)Supporting an API is a very important task, and Google Cloud provides three API management tools: Cloud Endpoints, API Gateway, and Apigee API Management.

[01:25](javascript:;)Cloud Endpoints is a distributed API management system that uses a distributed Extensible Service Proxy, which is a service proxy that runs in its own Docker container.

[01:35](javascript:;)The goal is to help you create and maintain even the most demanding APIs with low latency and high performance.

[01:43](javascript:;)Cloud Endpoints provides an API console, hosting, logging, monitoring, and other features to help you create, share, maintain, and secure your APIs.

[01:55](javascript:;)You can use Cloud Endpoints with any APIs that support the OpenAPI Specification.

[02:00](javascript:;)Cloud Endpoints supports applications running in App Engine, Google Kubernetes Engine, and Compute Engine.

[02:08](javascript:;)Clients include Android, iOS, and Javascript.

[02:09](javascript:;)API Gateway is another API management tool.

[02:15](javascript:;)Web-based services today provide a huge variety of functionality, meaning everything from map, weather, and image services, to games, auctions, and many other service types.

[02:26](javascript:;)Service providers have many options for how to implement, deploy, and manage their services.

[02:33](javascript:;)For example, one service might be developed in Java or .

[02:38](javascript:;)NET, while another uses Node.js.

[02:41](javascript:;)Backend implementations can also vary for a single service provider.

[02:45](javascript:;)A service provider might have legacy services implemented using one architecture, and new services implemented using a completely different architecture.

[02:53](javascript:;)API Gateway enables you to provide secure access to your backend services through a well-defined REST API that is consistent across all of your services, regardless of the service implementation.

[03:06](javascript:;)Clients consume your REST APIS to implement standalone apps for a mobile device or tablet, through apps running

[03:12](javascript:;)in a browser, or through any other type of app that can make a request to an HTTP endpoint.

[03:19](javascript:;)Another Google Cloud platform available for developing and managing API proxies is Apigee API Management.

[03:26](javascript:;)Unlike Cloud Endpoints, Apigee API Management has a specific focus on business problems, like rate limiting, quotas, and analytics.

[03:34](javascript:;)In fact, many Apigee API Management users provide a software service to other companies.

[03:41](javascript:;)Backend services for Apigee API Management don't have to be in Google Cloud, and as a result, engineers also often use it to take apart legacy applications.

[03:49](javascript:;)So, instead of replacing a large, important application in one move, they can use Apigee API Management to peel off its services individually instead.

[03:58](javascript:;)This allows them to stand up microservices to implement each in turn until the legacy application can finally be retired.

Cloud Run-

[00:00](javascript:;)The final application platform we’ll explore in this section of the course is Cloud Run,

[00:04](javascript:;)a managed compute platform that lets you run stateless containers via web requests or Pub/Sub events.

[00:11](javascript:;)Cloud Run is serverless.

[00:14](javascript:;)That means it removes all infrastructure management tasks so you can focus on developing applications.

[00:19](javascript:;)It’s built on Knative, an open API and runtime environment built on Kubernetes that gives you freedom to move your workloads across different environments and platforms.

[00:29](javascript:;)It can be fully managed on Google Cloud, on Google Kubernetes Engine, or anywhere Knative runs.

[00:36](javascript:;)Cloud Run is fast.

[00:38](javascript:;)It can automatically scale up and down from zero almost instantaneously, and it charges you only for the

[00:44](javascript:;)resources you use, calculated down to the nearest 100 milliseconds, so you‘ll never pay for your over-provisioned resources.

[00:52](javascript:;)The Cloud Run developer workflow is a straightforward three-step process.

[00:58](javascript:;)First, you write your application using your favorite programming language.

[01:03](javascript:;)This application should start a server that listens for web requests.

[01:07](javascript:;)Second, you build and package your application into a container image.

[01:12](javascript:;)Third, the container image is pushed to Artifact Registry, where Cloud Run will deploy it.

[01:19](javascript:;)Note that Cloud Run can only deploy images that are stored in Artifact Registry.

[01:25](javascript:;)You can build, push and deploy your own code from your local source if you have the required permissions.

[01:32](javascript:;)You can also deploy an image that already exists in Artifact Registry.

[01:37](javascript:;)Once you’ve deployed your container image, you’ll get a unique HTTPS URL back.

[01:42](javascript:;)Cloud Run then starts your container on demand to handle requests, and ensures that all incoming requests are handled by dynamically adding and removing containers.

[01:53](javascript:;)Because Cloud Run is serverless, it means that you, as a developer, can focus on building your application and not on building and maintaining the infrastructure that powers it.

[02:03](javascript:;)For some use cases, a container-based workflow is great, because it gives you a great amount of transparency and flexibility.

[02:10](javascript:;)If you build the container image you have the power to decide exactly what file ends up in your container image, and how it gets there.

[02:18](javascript:;)However, building an application is hard enough already, let alone having to think about containerization and the responsibilities that come with that.

[02:26](javascript:;)Sometimes, you’re just looking for a way to turn source code into an HTTPS endpoint, and you

[02:31](javascript:;)want your vendor to make sure your container image is secure, well-configured built in a consistent way.

[02:38](javascript:;)With Cloud Run, you can do both.

[02:42](javascript:;)You can use a container-based workflow, as well as a source-based workflow.

[02:47](javascript:;)If you use the source-based approach, you’ll deploy your source code, instead of a container image.

[02:53](javascript:;)Cloud Run then builds your source and packages the application into a container image for you.

[02:59](javascript:;)Cloud Run does this using Buildpacks - an open source project.

[03:04](javascript:;)Cloud Run handles HTTPS serving for you.

[03:07](javascript:;)That means you only have to worry about handling web requests, and you can let Cloud Run take care of adding the encryption.

[03:14](javascript:;)By default, your application is exposed on a unique subdomain of the global run.app domain.

[03:22](javascript:;)You can also use your own, custom domain.

[03:25](javascript:;)Cloud Run manages everything else: Generating a valid SSL certificate Configuring SSL termination correctly with secure

[03:32](javascript:;)settings And handling incoming requests, decrypting them, and forwarding them to your application The pricing model on

[03:39](javascript:;)Cloud Run is unique; as you only pay for the system resources you use while a

[03:44](javascript:;)container is handling web requests, with a granularity of 100ms, and when it’s starting or shutting down.

[03:53](javascript:;)You don’t pay for anything if your container doesn’t handle requests.

[03:57](javascript:;)Additionally, there is a small fee for every one million requests you serve.

[04:02](javascript:;)The price of container time increases With CPU and memory.

[04:07](javascript:;)A container with more vCPU and memory is more expensive.

[04:11](javascript:;)Today, Cloud run can allocate up to 4 vCPUs and 8 gigabytes of memory.

[04:17](javascript:;)Most of the other compute products (such as Compute Engine), charge for servers as long as they’re running, even if you’re not using them.

[04:23](javascript:;)That means you’re often paying for idle server capacity.

[04:28](javascript:;)You can use Cloud Run to run any binary, as long as it’s compiled for Linux sixty-four bit.

[04:34](javascript:;)Now, this means you can use Cloud Run to run web applications written using popular languages, such as: Java Python Node.js PHP Go C++ And

[04:45](javascript:;)you can also run code written in less popular languages: Cobol Haskell Perl As long as your app handles web requests, you’re good to go.

[Cloud Run](https://cloud.google.com/run) is a managed compute platform that enables you to run stateless containers that are invocable via HTTP requests. Cloud Run is serverless: it abstracts away all infrastructure management, so you can focus on what matters most — building great applications.

[Cloud Run](https://cloud.google.com/run) is built from [Knative](https://cloud.google.com/knative/" \t "_blank), letting you choose to run your containers either fully managed with Cloud Run, or in your [Google Kubernetes Engine](https://cloud.google.com/kubernetes) cluster with Cloud Run on GKE.

**Google Cloud Shell** is a virtual machine that is loaded with development tools. It offers a persistent 5GB home directory and runs on the Google Cloud.

Google Cloud Shell provides command-line access to your Google Cloud resources.

**gcloud** is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab-completion.

Cloud Build is a service that executes your builds on GCP. It executes a series of build steps, where each build step is run in a Docker container to produce your application container (or other artifacts) and push it to Cloud Registry, all in one command.

Once pushed to the registry, you will see a SUCCESS message containing the image name (gcr.io/[PROJECT-ID]/helloworld). The image is stored in Artifact Registry and can be re-used if desired.

Hello Cloud Run [APPRUN]- Lab Completed

Quiz: Applications in the Cloud-

Cloud Run can only pull images from:

Artifact Registry

App Engine is best suited to the development and hosting of which type of application?

A web application

Which Google Cloud service should you choose to perform business analytics and billing on a customer-facing API?

Apigee API Management

Select the managed compute platform that lets you run stateless containers through web requests or Pub/Sub events.

Cloud Run

What are the advantages of using App Engine’s flexible environment instead of its standard environment? (Select 3).  
You can install third-party binaries across all languages.  
Your application can write to the local disk.

You can use SSH to connect to the virtual machines on which your application runs.

Which statements are true about App Engine? (Select 2).

The daily billing for an App Engine application can drop to zero.

App Engine manages the hardware and networking infrastructure required to run your code.

DEVELOPING AND DEPLOYING IN THE CLOUD:

Development in the Cloud-

[00:00](javascript:;)Many users develop impressive applications using Google Cloud’s products and services.

[00:05](javascript:;)And when an app is ready, Google Cloud can also be used to deploy it.

[00:10](javascript:;)In this section of the course, we’ll explore Google Cloud methods for development in the cloud, which includes Cloud Source Repositories, Cloud Functions, and Terraform.

[00:19](javascript:;)After that, we’ll look at deployment with infrastructure as code.

[00:23](javascript:;)Let’s begin by looking at development in the cloud.

[00:27](javascript:;)Many Google Cloud customers use Git repositories to store, version, and manage their source code trees.

[00:33](javascript:;)That means they either run their own Git instances, which is a great option if total control

[00:37](javascript:;)is required, or they use a hosted-Git provider, which means less work if total control isn’t required.

[00:44](javascript:;)But what if there were a third option, where you could keep code private to a Google

[00:47](javascript:;)Cloud project and use IAM permissions to protect it, but not have to maintain the Git instance yourself?

[00:55](javascript:;)Well, that third option is available with Cloud Source Repositories.

[01:00](javascript:;)Cloud Source Repositories provides full-featured Git repositories hosted on Google Cloud that support the collaborative

[01:06](javascript:;)development of any application or service, including those that run on App Engine and Compute Engine.

[01:13](javascript:;)With Cloud Source Repositories, you can have any number of private Git repositories.

[01:18](javascript:;)This allows code associated with a cloud project to be organized the way you choose.

[01:24](javascript:;)It also allows Google Cloud diagnostics tools, like Debugger and Error Reporting, to use the code from

[01:29](javascript:;)Git repositories to track down issues to specific errors in deployed code without slowing down your users.

[01:36](javascript:;)If your code is already in GitHub or BitBucket repositories, it can be migrated into your cloud project and used just like any other repository, including browsing and diagnostics.

[01:46](javascript:;)Many applications contain event-driven parts.

[01:49](javascript:;)For example, maybe you have an application that lets users upload images.

[01:54](javascript:;)When that event takes place, the image might need to be processed in a few different

[01:57](javascript:;)ways, like converting a thumbnail into different sizes, and storing each new file in a repository.

[02:06](javascript:;)You could integrate this function into your application, but then you’d have to provide compute resources for it–whether it happens once a millisecond or once a day.

[02:15](javascript:;)With Cloud Functions, you could write a single-purpose function that completes the necessary image manipulations and then arrange for it to automatically run whenever a new image is uploaded.

[02:26](javascript:;)Cloud Functions is a lightweight, event-based, asynchronous compute solution that allows you to create small, single-purpose

[02:31](javascript:;)functions that respond to cloud events, without the need to manage a server or a runtime environment.

[02:39](javascript:;)You can use these functions to construct application workflows from individual business logic tasks.

[02:44](javascript:;)You can also use Cloud Functions to connect and extend cloud services.

[02:49](javascript:;)You’re billed to the nearest 100 milliseconds, but only while your code is running.

[02:54](javascript:;)Cloud Functions supports writing source code in a number of programming languages.

[02:59](javascript:;)These include Node.js, Python, Go, Java, .

[03:04](javascript:;)Net Core, Ruby, and PHP.

[03:08](javascript:;)For more information about the supported specific versions, refer to the runtimes documentation.

[03:15](javascript:;)Events from Cloud Storage and Pub/Sub can trigger Cloud Functions asynchronously, or you can use HTTP invocation for synchronous execution.

Deployment: Infrastructure as Code-

[00:00](javascript:;)Creating an environment in Google Cloud can mean lots of work–like setting up a compute network and storage resources and then keeping track of their configurations.

[00:08](javascript:;)This process can be done manually by writing the commands you need to set up your environment the way you want.

[00:15](javascript:;)However, this is labor-intensive and requires updating commands if you want to change the environment or manually writing new commands if you want to clone an environment.

[00:24](javascript:;)It’s more efficient to use a template.

[00:27](javascript:;)Using a template allows you to write the specifications for your application environment in the same way you’d write a configuration

[00:32](javascript:;)file, but your template can then be deployed in a scaled environment to quickly create as many identical application environments as needed.

[00:41](javascript:;)This can be done with Terraform.

[00:44](javascript:;)To use Terraform, you create a template file, using HashiCorp Configuration Language (HCL), that describes what you want the components of your environment to look like.

[00:54](javascript:;)Terraform then uses that template to determine the actions needed to create the environment your template describes.

[01:01](javascript:;)If you need to change the environment, you can edit your template and then use Terraform to update the environment to match the change.

[01:09](javascript:;)You can store and version-control your Terraform templates in Cloud Source Repositories.

**Terraform** enables you to safely and predictably create, change, and improve infrastructure. It is an open-source tool that codifies APIs into declarative configuration files that can be shared among team members, treated as code, edited, reviewed, and versioned.

Automating the Deployment of Infrastructure using Terraform- Lab Completed

Quiz: Developing and Deploying in the Cloud-

Why might a Google Cloud customer choose to use Cloud Functions?

Their application contains event-driven code that they don't want to provision compute resources for.

Select the advantage of putting the event-driven components of your application into Cloud Functions.

Cloud Functions handles scaling these components seamlessly.

Why would a developer choose to store source code in Cloud Source Repositories? (Select 2)  
To reduce work

To keep code private to a Google Cloud project

Why might a Google Cloud customer choose to use Terraform?

Terraform can be used as an infrastructure management system for Google Cloud resources.

LOGGING AND MONITORING IN THE CLOUD:

The Importance of Monitoring-

[00:00](javascript:;)In this section of the course we’ll transition our focus from developing and deploying in the cloud, to logging and monitoring.

[00:06](javascript:;)Let’s begin with monitoring.

[00:10](javascript:;)Monitoring is the foundation of product reliability.

[00:13](javascript:;)It reveals what needs urgent attention and shows trends in application usage patterns, which can

[00:18](javascript:;)yield better capacity planning and generally help improve an application client's experience and lessen their pain.

[00:26](javascript:;)In Google's Site Reliability Engineering book, which is available to read at landing.google.com/sre/books, monitoring is defined as: "Collecting, processing, aggregating, and

[00:38](javascript:;)displaying real-time quantitative data about a system, such as query counts and types, error counts and types, processing times, and server lifetimes."

[00:50](javascript:;)With monitoring, you can ensure continued system operations, uncover trend analyses over time, build dashboards, alert personnel when systems violate predefined

[00:59](javascript:;)service level objectives (SLOs), compare systems and systems changed, and provide data for improved incident response–just to name a few tasks.

[01:11](javascript:;)An application client normally only sees the public side of a product, and as a result, developers and business stakeholders both tend to think

[01:19](javascript:;)that the most crucial way to make the client happy is by spending the most time and effort on developing that part of the product.

[01:26](javascript:;)However, to be truly reliable, even the very best products still must be deployed into environments with enough capacity to handle the anticipated client load.

[01:36](javascript:;)Great products also need thorough testing, preferably automated testing, and a refined continuous integration/continuous development (CI/CD) release pipeline.

[01:45](javascript:;)Postmortems and root cause analyses are the DevOps team's way of letting the client know why an incident happened and why it’s unlikely to happen again.

[01:54](javascript:;)In this context we’re discussing a system or software failure, but the term “incident” can also be used to describe a breach of security.

[02:01](javascript:;)Here, transparency is key to building trust.

[02:02](javascript:;)why an incident happened and why it’s unlikely to happen again.

[02:03](javascript:;)In this context we’re discussing a system or software failure, but the term “incident” can also be used to describe a breach of security.

[02:05](javascript:;)Here, transparency is key to building trust.

[02:06](javascript:;)We need our products to improve continually, and we need data we can receive from monitoring to make sure that happens.

[02:13](javascript:;)We need dashboards to provide business intelligence so our DevOps personnel have the data they need to do their jobs.

[02:20](javascript:;)We need automated alerts because humans tend to look at things only when there's something important to look at.

[02:26](javascript:;)An even better option is to construct automated systems to handle as many alerts as possible so humans only have to look at the most crucial issues.

[02:34](javascript:;)Finally, we need monitoring tools that help provide data crucial to debugging application functional and performance issues.

[02:43](javascript:;)We’ll look more closely at Google’s integrated monitoring tools a bit later in this module.

Measuring Performance and Reliability-

[00:00](javascript:;)There are “Four Golden Signals” that measure a system’s performance and reliability.

[00:04](javascript:;)They are latency, traffic, saturation, and errors.

[00:11](javascript:;)Latency measures how long it takes a particular part of a system to return a result.

[00:16](javascript:;)Latency is important because: It directly affects the user experience.

[00:22](javascript:;)Changes in latency could indicate emerging issues.

[00:26](javascript:;)Its values may be tied to capacity demands.

[00:28](javascript:;)And it can be used to measure system improvements.

[00:32](javascript:;)But how exactly is it measured?

[00:36](javascript:;)Sample latency metrics include: Page load latency Number of requests waiting for a thread Query duration Service response time Transaction duration Time

[00:51](javascript:;)to first response Time to complete data return The next signal is traffic, which measures how many requests are reaching your system.

[01:02](javascript:;)Traffic is important because: It’s an indicator of current system demand.

[01:06](javascript:;)Its historical trends are used for capacity planning.

[01:09](javascript:;)And it’s a core measure when calculating infrastructure spend.

[01:15](javascript:;)Sample traffic metrics include: number of HTTP requests per second number of requests for static vs.

[01:23](javascript:;)dynamic content Network I/O number of concurrent sessions number of transactions per second number of retrievals

[01:34](javascript:;)per second number of active requests number of write ops number of read ops And number

[01:43](javascript:;)of active connections The third signal is saturation, which measures how close to capacity a system is.

[01:52](javascript:;)It’s important to note, though, that capacity is often a subjective measure that depends on the underlying service or application.

[02:00](javascript:;)Saturation is important because: It's an indicator of how full the service is.

[02:05](javascript:;)It focuses on the most constrained resources.

[02:07](javascript:;)And it’s frequently tied to degrading performance as capacity is reached.

[02:14](javascript:;)Sample capacity metrics include: % memory utilization % thread pool utilization % cache utilization % disk utilization % CPU utilization Disk quota Memory quota number of

[02:31](javascript:;)of available connections And number of of users on the system The fourth signal is errors, which are events that measure system failures or other issues.

[02:44](javascript:;)Errors are often raised when a flaw, failure, or fault in a computer program or system causes it to produce incorrect or unexpected results, or behave in unintended ways.

[02:55](javascript:;)Errors are important because: They may indicate that something is failing.

[02:59](javascript:;)They may indicate configuration or capacity issues.

[03:03](javascript:;)They can indicate service level objective violations.

[03:06](javascript:;)And an error might mean it's time to send out an alert.

[03:11](javascript:;)Sample error metrics include: Wrong answers or incorrect content number of 400/500 HTTP codes number of failed

[03:21](javascript:;)requests number of exceptions number of stack traces Servers that fail liveness checks And number of dropped connections

Understanding SLIs, SLOs and SLAs-

[00:00](javascript:;)Now let’s shift our focus to SLIs, SLOs and SLAs, which are all types of targets set for a system’s Four Golden Signal metrics.

[00:10](javascript:;)Service level indicators, or SLIs, are carefully selected monitoring metrics that measure one aspect of a service's reliability.

[00:18](javascript:;)Ideally, SLIs should have a close linear relationship with your users' experience of that reliability, and we recommend expressing

[00:26](javascript:;)them as the ratio of two numbers: the number of good events divided by the count of all valid events.

[00:33](javascript:;)A Service level objective, or SLO, combines a service level indicator with a target reliability.

[00:41](javascript:;)If you express your SLIs as is commonly recommended, your SLOs will generally be somewhere just short of 100%, for example, 99.9%, or "three nines."

[00:54](javascript:;)You can't measure everything, so when possible, you should choose SLOs that are S.M.A.R.T. SLOs should be specific.

[01:01](javascript:;)A question such as “Is the site fast enough for you?” is not specific; it's subjective.

[01:07](javascript:;)A statement such as “The 95th percentile of results are returned in under 100 milliseconds” is specific.

[01:14](javascript:;)SLOs need to be based on indicators that are measurable.

[01:18](javascript:;)A lot of monitoring is numbers, grouped over time, with math applied.

[01:22](javascript:;)An SLI must be a number or a delta; something we can measure and place in a mathematical equation.

[01:30](javascript:;)SLO goals should be achievable.

[01:31](javascript:;)"100% Availability" might sound good, but it's not possible to obtain, let alone maintain, over an extended window of time.

[01:42](javascript:;)SLOs should be relevant.

[01:43](javascript:;)Does it matter to the user?

[01:46](javascript:;)Will it help achieve application-related goals?

[01:49](javascript:;)If not, then it’s a poor metric.

[01:52](javascript:;)And SLOs should be time-bound.

[01:54](javascript:;)You want a service to be 99% available?

[01:57](javascript:;)That’s fine.

[01:58](javascript:;)Is that per year?

[01:59](javascript:;)Per month?

[02:00](javascript:;)Per day?

[02:01](javascript:;)Does the calculation look at specific windows of set time, from Sunday to Sunday for example, or is it a rolling period of the last seven days?

[02:09](javascript:;)If we don't know the answers to those types of questions, it can’t be measured accurately.

[02:15](javascript:;)And then there are Service Level Agreements, or SLAs, which are commitments made to your customers that your systems and applications will have only a certain amount of “down time.”

[02:26](javascript:;)An SLA describes the minimum levels of service that you promise to provide to your customers and what happens when you break that promise.

[02:34](javascript:;)If your service has paying customers, an SLA may include some way of compensating them with

[02:38](javascript:;)refunds or credits when that service has an outage that is longer than this agreement allows.

[02:45](javascript:;)To give you the opportunity to detect problems and take remedial action before your reputation is damaged,

[02:50](javascript:;)your alerting thresholds are often substantially higher than the minimum levels of service documented in your SLA.

[02:59](javascript:;)For SLOs, SLIs, and SLAs to help improve service reliability, all parts of the business must agree that they are

[03:05](javascript:;)an accurate measure of user experience and must also agree to use them as a primary driver for decision making.

[03:13](javascript:;)Being out of SLO must have concrete, well-documented consequences, just as there are consequences for breaching SLAs.

[03:22](javascript:;)For example, slowing down the rate of change and directing more engineering effort toward eliminating risks and

[03:27](javascript:;)improving reliability are actions that could be taken to get your product back to meeting its SLOs faster.

[03:35](javascript:;)Operations teams need strong executive support to enforce these consequences and effect change in your development practice.

Integrated Observability Tools-

[00:00](javascript:;)Let’s wrap up this section of the course by taking a look at Google Cloud’s integrated monitoring, logging, error reporting, and debugging tools.

[00:09](javascript:;)If you've ever worked with on-premises environments, you know that you can physically touch the servers.

[00:15](javascript:;)If an application becomes unresponsive, someone can physically determine why that happened.

[00:20](javascript:;)In the cloud though, the servers aren't yours—they're Google’s—and you can’t physically inspect them.

[00:27](javascript:;)So the question becomes, how do you know what's happening with your server, or database, or application?

[00:34](javascript:;)The answer is by using Google’s integrated observability tools.

[00:39](javascript:;)Observability starts with signals, which are metric, logging, and trace data captured and integrated into Google products from the hardware layer up.

[00:49](javascript:;)From those products: The signal data flows into the Google Cloud operation's tools where it can be visualized in dashboards and through the Metrics Explorer.

[00:59](javascript:;)Automated and custom logs can be dissected and analyzed in the Logs Explorer.

[01:04](javascript:;)Services can be monitored for compliance with service level objectives (SLOs), and error budgets can be tracked.

[01:11](javascript:;)Health checks can be used to check uptime and latency for external-facing sites and services.

[01:16](javascript:;)And running applications can be debugged and profiled.

[01:21](javascript:;)When incidents occur: Signal data can generate automated alerts to code or, through various information channels, to key personnel.

[01:30](javascript:;)Error Reporting can help operations and developer teams spot, count, and analyze crashes in cloud-based services.

[01:38](javascript:;)Service Level Objectives should be adhered to.

[01:40](javascript:;)And the visualization and analysis tools can then help troubleshoot what's happening in Google Cloud.

[01:47](javascript:;)Ultimately, you won't miss that easy server access, because Google provides more precise insights into your Cloud install than you ever had on-premises.

[01:57](javascript:;)Over the next few videos we’ll explore the products and tools offered by Google Cloud that

[02:01](javascript:;)are most applicable for those in operations roles that work with monitoring, logging, error reporting, and debugging.

Monitoring Tools-

[00:00](javascript:;)When DevOps personnel want to track exactly what's happening inside Google Cloud projects, they often first think of monitoring.

[00:09](javascript:;)As we stated previously, monitoring starts with signal data.

[00:13](javascript:;)Metrics take measurements and use math to align those measurements over time.

[00:18](javascript:;)For example, it might be taking raw CPU usage measurement values and averaging them to produce a single value per minute.

[00:27](javascript:;)Google Cloud, by default, collects more than a thousand different streams of metric data, which can be incorporated into dashboards, alerts, and several other key tools.

[00:38](javascript:;)When data scientists run massive, scalable queries in BigQuery, it’s important for them to know how many queries are

[00:44](javascript:;)currently in flight, how many bytes have been scanned and added to the bill, and data slot usage patterns.

[00:52](javascript:;)It could also be critical to DevOps teams running containerized applications in Cloud Run to know CPU and memory utilization and app bill time.

[01:02](javascript:;)If those same DevOps teams want to augment the signal metrics from their custom application wherever it's running, they could use the open-source OpenTelementry and create their own metrics.

[01:14](javascript:;)And workloads on Compute Engine will benefit from CPU and memory utilization data, along with uptime, disk throughput, and many other metrics.

[01:25](javascript:;)Cloud Monitoring provides visibility into the performance, uptime, and overall health of cloud-powered applications.

[01:33](javascript:;)It collects metrics, events, and metadata from projects, logs, services, systems, agents, custom code, and various common application components, including Cassandra, Nginx, Apache Web Server, and Elasticsearch.

[01:52](javascript:;)Cloud Monitoring ingests that data and generates insights via dashboards, Metrics Explorer charts, and automated alerts.

[01:58](javascript:;)In this video we’ll take a look at Google Cloud’s integrated logging tools.

Logging Tools-

[00:00](javascript:;)In this video we’ll take a look at Google Cloud’s integrated logging tools.

[00:05](javascript:;)Cloud Logging allows users to collect, store, search, analyze, monitor, and alert on log entries and events.

[00:14](javascript:;)Automated logging is integrated into Google Cloud products like App Engine, Cloud Run, Compute Engine VMs running the logging agent, and GKE.

[00:25](javascript:;)Most log analysis starts with Google Cloud’s integrated Logs Explorer.

[00:30](javascript:;)Logging entries can also be exported to several destinations for alternative or further analysis.

[00:36](javascript:;)Pub/Sub messages can be analyzed in near-real time using custom code or stream processing technologies like Dataflow.

[00:45](javascript:;)BigQuery allows analysts to examine logging data through SQL queries.

[00:50](javascript:;)And archived log files in Cloud Storage can be analyzed with several tools and techniques.

[00:56](javascript:;)Log data can be exported as files to Cloud Storage, as messages through Pub/Sub, or into BigQuery tables.

[01:04](javascript:;)Log-based metrics can be created and integrated into Cloud Monitoring dashboards, alerts, and service SLOs.

[01:12](javascript:;)Default log retention in Cloud Logging depends on the log type.

[01:16](javascript:;)Data access logs are retained by default for 30 days, but this is configurable up to a maximum of 3,650 days.

[01:25](javascript:;)Admin logs are stored by default for 400 days.

[01:29](javascript:;)Logs can be exported to Cloud Storage or BigQuery to extend retention.

[01:34](javascript:;)The logs visible to you in Cloud Logging vary, depending on which Google Cloud resources you're using in your project or organization.

[01:42](javascript:;)Four key log categories are audit logs, agent logs, network logs, and service logs.

[01:51](javascript:;)Cloud Audit Logs helps answer the question, "Who did what, where, and when?"

[01:57](javascript:;)Admin activity tracks configuration changes.

[02:01](javascript:;)Data access tracks calls that read the configuration or metadata of resources and user-driven calls that create, modify, or read user-provided resource data.

[02:13](javascript:;)System events are non-human Google Cloud administrative actions that change the configuration of resources.

[02:21](javascript:;)And Access Transparency provides you with logs that capture the actions Google personnel take when accessing your content.

[02:29](javascript:;)Agent logs use a Google-customized and packaged Fluentd agent that can be installed on any AWS or Google Cloud

[02:35](javascript:;)VM to ingest log data from Google Cloud instances–for example, Compute Engine, Managed VMs, or Containers–and AWS EC2 instances.

[02:48](javascript:;)Network logs provide both network and security operations with in-depth network service telemetry.

[02:55](javascript:;)VPC Flow Logs records samples of VPC network flow and can be used for network monitoring, forensics, real-time security analysis, and expense optimization.

[03:09](javascript:;)Firewall Rules Logging allows you to audit, verify, and analyze the effects of your firewall rules.

[03:16](javascript:;)NAT Gateway logs capture information on NAT network connections and errors.

[03:23](javascript:;)Service logs provide access to logs created by developers deploying code to Google Cloud.

[03:29](javascript:;)For example, if they build a container using Node.js and deploy it to Cloud Run, any logging

[03:35](javascript:;)to Standard Out or Standard Error will automatically be sent to Cloud Logging for easy, centralized viewing.

[03:41](javascript:;)Let’s round off this section of the course by taking a look at the tools offered by Google Cloud for error reporting and debugging.

Error Reporting and Debugging Tools-

[00:00](javascript:;)Let’s round off this section of the course by taking a look at the tools offered by Google Cloud for error reporting and debugging.

[00:07](javascript:;)Error Reporting counts, analyzes, and aggregates the crashes in your running cloud services.

[00:15](javascript:;)Crashes in most modern languages are “Exceptions,” which aren’t caught and handled by the code itself.

[00:21](javascript:;)Its management interface displays the results with sorting and filtering capabilities.

[00:26](javascript:;)A dedicated view shows the error details: time chart, occurrences, affected user count, first- and last-seen dates, and a cleaned exception stack trace.

[00:38](javascript:;)You can also create alerts to receive notifications on new errors.

[00:39](javascript:;)Cloud Trace, based on the tools Google uses on its production services, is a tracing system

[00:44](javascript:;)that collects latency data from your distributed applications and displays it in the Google Cloud console.

[00:51](javascript:;)Trace can capture traces from applications deployed on App Engine, Compute Engine VMs, and Google Kubernetes Engine containers.

[01:00](javascript:;)Performance insights are provided in near-real time, and Trace automatically analyzes all of your application's traces to generate in-depth latency reports to surface performance degradations.

[01:13](javascript:;)Trace continuously gathers and analyzes trace data to automatically identify recent changes to your application's performance.

[01:21](javascript:;)Poorly performing code increases the latency and cost of applications and web services every day, without anyone knowing or doing anything about it.

[01:31](javascript:;)Cloud Profiler changes this by using statistical techniques and extremely low-impact instrumentation that runs across all production

[01:38](javascript:;)application instances to provide a complete CPU and heap picture of an application without slowing it down.

[01:47](javascript:;)With broad platform support that includes Compute Engine VMs, App Engine, and Kubernetes, it allows developers to analyze

[01:55](javascript:;)applications running anywhere, including Google Cloud, other cloud platforms, or on-premises, with support for Java, Go, Python, and Node.js.

[02:08](javascript:;)Cloud Profiler presents the call hierarchy and resource consumption of the relevant function in an interactive flame graph that

[02:14](javascript:;)helps developers understand which paths consume the most resources and the different ways in which their code is actually called.

Quiz: Logging and Monitoring in the Cloud-

Which option describes a commitment made to your customers that your systems and applications will have only a certain amount of “downtime”?

Service level agreement

Which definition best describes a service level indicator (SLI)?

A carefully selected monitoring metric that measures one aspect of a service's reliability.

Select the two correct statements about Cloud Logging.

Cloud Logging lets you define metrics based on your logs.

Cloud Logging lets you view logs from your applications and filter and search on them.

You want to create alerts on your Google Cloud resources, such as when health checks fail. Which is the best Google Cloud product to use?

Cloud Monitoring

There are “Four Golden Signals” that measure a system’s performance and reliability. What are they?

Latency, traffic, saturation, errors

COURSE SUMMARY:

[00:00](javascript:;)Congratulations on completing the Google Cloud Core Infrastructure course!

[00:04](javascript:;)Before you go, let’s take a few minutes to review what we’ve covered.

[00:10](javascript:;)In module 1, you were introduced to Google Cloud and cloud computing.

[00:14](javascript:;)Specifically, you explored: The concept of managed infrastructure and managed services, through IaaS, or infrastructure as a service, and PaaS, or platform as a service.

[00:24](javascript:;)The Google Cloud network.

[00:27](javascript:;)Google Cloud’s focus on security throughout our infrastructure.

[00:30](javascript:;)How Google publishes key elements of technology using open source licenses.

[00:35](javascript:;)And Google Cloud’s pricing structure and billing tools.

[00:39](javascript:;)In module 2, you learned about the Google Cloud Resource Hierarchy, which is made up of four levels: resources, projects, folders, and an organization node.

[00:50](javascript:;)You also learned about: Defining policies and their downward inheritance.

[00:54](javascript:;)When to use Identity and Access Management, or IAM, And the four ways to access and interact with Google

[01:00](javascript:;)Cloud: through the Google Cloud console, the Cloud SDK and Cloud Shell, APIs, and the Google Cloud console Mobile App.

[01:09](javascript:;)In module 3, you explored how Compute Engine works, with a focus on virtual machines and virtual networking.

[01:16](javascript:;)You were introduced to: The VPC, or virtual private cloud.

[01:21](javascript:;)Compute Engine’s Autoscaling feature.

[01:24](javascript:;)And important Google Virtual Private Cloud compatibility features, like routing tables, firewalls, VPC peering and shared VPC, all of which result in the need for less network management.

[01:37](javascript:;)You also explored Cloud Load Balancing, a fully distributed, software-defined, managed service for all your traffic.

[01:44](javascript:;)Finally, you compared how on-premises or other-cloud networks can be interconnected with a Google VPC.

[01:51](javascript:;)In module 4, you explored Google Cloud's five core storage options: Cloud Storage, Cloud Bigtable, Cloud SQL, Cloud Spanner, and Firestore.

[02:03](javascript:;)You were also examined the four storage classes that make up Cloud Storage: \* Standard Storage, which is used for frequently

[02:08](javascript:;)accessed hot data, \* Nearline Storage and Coldline Storage, which are used for less-frequently accessed cool data, \* and Archive Storage.

[02:18](javascript:;)In module 5, you learned about containers, which are invisible boxes around your code and its dependencies.

[02:25](javascript:;)You were introduced to three container-based products: Kubernetes, an open-source platform for managing containerized workloads and services.

[02:34](javascript:;)Google Kubernetes Engine (GKE), a Google-hosted managed Kubernetes service in the cloud.

[02:40](javascript:;)And Anthos, Google’s answer to modern hybrid and multi-cloud distributed systems and services management.

[02:47](javascript:;)In module 6, the focus was on developing applications in the cloud.

[02:51](javascript:;)You explored: App Engine, a fully managed, serverless platform for developing and hosting web applications at scale, and the two of App Engine environments: standard and flexible.

[03:01](javascript:;)Three API management tools provided by Google Cloud: Cloud Endpoints, API Gateway, and Apigee API Management.

[03:09](javascript:;)And Cloud Run, a managed compute platform that lets you run stateless containers via web requests or Pub/Sub events.

[03:17](javascript:;)The focus for module 7 was developing and deploying in the cloud.

[03:21](javascript:;)You learned about: Cloud Source Repositories, which are full-featured Git repositories hosted on Google Cloud.

[03:27](javascript:;)Cloud Functions, a lightweight, event-based, asynchronous compute solution to create single-purpose functions.

[03:33](javascript:;)And Terraform, which lets you use a template to write the specifications for your application environment in the same way you’d write a configuration file.

[03:42](javascript:;)And in the final module, you focused on logging and monitoring on Google Cloud.

[03:47](javascript:;)The “Four Golden Signals” that measure a system’s performance and reliability: latency, traffic, saturation, and errors.

[03:55](javascript:;)Service level indicators (SLIs), service level objectives (SLOs), and service level agreements (SLAs), which are all types of targets set for a system’s Four Golden Signal metrics.

[04:06](javascript:;)And finally, Google’s integrated observability tools, which include Cloud Monitoring, Cloud Logging, Error Reporting, Cloud Trace, and Cloud Profiler.

[04:15](javascript:;)We hope that this course is just the beginning of your Google Cloud journey.

[04:21](javascript:;)For more training and hands-on practice, explore the different learning paths available at cloud.google.com/training.

[04:29](javascript:;)And if you’re interested in validating your expertise and showcasing your ability to transform businesses with Google Cloud technology, you might consider working toward a Google Cloud certification.

[04:39](javascript:;)You can learn more about Google Cloud’s certification offerings at cloud.google.com/certification.

[04:46](javascript:;)Thanks for completing this course.

[04:48](javascript:;)We’ll see you next time!

=======================================================================

am sending you the workaround for the lab Automating the Deployment of Infrastructure Using Terraform. Attempt the lab using the following workaround and let me know. LAB: Automating the Deployment of Infrastructure Using Terraform Task 1. Set up Terraform and Cloud Shell In the Cloud Console, click Activate Cloud Shell (). If prompted, click Continue. To confirm that Terraform is installed. Command:      terraform --version   3. To create a directory for your Terraform configuration, run the following Command:      mkdir tfinfra       4. In Cloud Shell, click Open editor (). Note: If you see the message "Unable to load code editor because third-party cookies are disabled", click Open in New Window. In the left pane of the code editor, expand the tfinfra folder. To create a new file inside tfinfra folder, right-click on tfinfra folder and then click New File. Name the new file [provider.tf](http://provider.tf/), and then open it. Copy the code below into [provider.tf](http://provider.tf/):      provider "google" {}       8.   To save [provider.tf](http://provider.tf/), click File > Save.       9.   To initialize Terraform in Cloud Shell. Command:      cd tfinfra      terraform init

Zoyeb says:Task 2. Create mynetwork and its resources Configure mynetwork To create a new file inside tfinfra, right-click on tfinfra folder and then click New File. Name the new file [mynetwork.tf](http://mynetwork.tf/), and then open it. Copy the below base code into [mynetwork.tf](http://mynetwork.tf/): # Create the mynetwork network resource [RESOURCE\_TYPE] "mynetwork" { name = [RESOURCE\_NAME]      # RESOURCE properties go here      } In [mynetwork.tf](http://mynetwork.tf/), replace [RESOURCE\_TYPE] with "google\_compute\_network" (with the quotes) and replace [RESOURCE\_NAME] with "mynetwork" (with the quotes). Add the following property to [mynetwork.tf](http://mynetwork.tf/): auto\_create\_subnetworks = "true"       6.   Verify that [mynetwork.tf](http://mynetwork.tf/) file look like this: # Create the mynetwork network resource "google\_compute\_network" "mynetwork" { name = "mynetwork" # RESOURCE properties go here auto\_create\_subnetworks = "true" }      7. To save [mynetwork.tf](http://mynetwork.tf/), click File > Save. Configure the firewall rule Add the following base code to [mynetwork.tf](http://mynetwork.tf/): # Add a firewall rule to allow HTTP, SSH, RDP and ICMP traffic on   mynetwork     resource [RESOURCE\_TYPE] "mynetwork-allow-http-ssh-rdp-icmp" {     name = [RESOURCE\_NAME]     # RESOURCE properties go here     } In [mynetwork.tf](http://mynetwork.tf/), replace [RESOURCE\_TYPE] with "google\_compute\_firewall" (with the quotes) and replace [RESOURCE\_NAME] with "mynetwork-allow-http-ssh-rdp-icmp" (with the quotes). Add the following property to [mynetwork.tf](http://mynetwork.tf/):      network = google\_compute\_network.mynetwork.self\_link       4.   Add the following properties to [mynetwork.tf](http://mynetwork.tf/):      allow {        protocol = "tcp"        ports    = ["22", "80", "3389"]      }      allow {        protocol = "icmp"      }      source\_ranges = ["0.0.0.0/0"]       5.   Verify that your [mynetwork.tf](http://mynetwork.tf/) file look like this: 6.  To save [mynetwork.tf](http://mynetwork.tf/), click File > Save. Configure the VM instance To create a new folder inside tfinfra, select the tfinfra folder, and then click File > New Folder. Name the new folder instance. To create a new file inside instance, right-click on instance folder and then click New File. Name the new file [main.tf](http://main.tf/), and then open it. You should have the following folder structure in Cloud Shell:     4.   Copy the following base code into [main.tf](http://main.tf/): resource [RESOURCE\_TYPE] "vm\_instance" {  name = [RESOURCE\_NAME]  # RESOURCE properties go here }     5.  In [main.tf](http://main.tf/), replace [RESOURCE\_TYPE] with "google\_compute\_instance" (with the quotes) and in [main.tf](http://main.tf/), replace [RESOURCE\_NAME] with "${var.instance\_name}" (with the quotes).     6.  Add the following properties to [main.tf](http://main.tf/):    zone         = "${var.instance\_zone}"    machine\_type = "${var.instance\_type}"  7. Add the following properties to [main.tf](http://main.tf/): boot\_disk {    initialize\_params {      image = "debian-cloud/debian-11"      }  }    8. Also, add the following properties to [main.tf](http://main.tf/):    network\_interface {      network = "${var.instance\_network}"      access\_config {      # Allocate a one-to-one NAT IP to the instance    }  }  9. Verify that [main.tf](http://main.tf/) looks like this:

Zoyeb says:9. Verify that [main.tf](http://main.tf/) looks like this: 10.  To save [main.tf](http://main.tf/), click File > Save. 11.   To create a new file inside instance, right-click on instance folder and then click New  File. Name the new file [variables.tf](http://variables.tf/), and then open it. 12.   Define the 4 input variables in [variables.tf](http://variables.tf/).    variable "instance\_name" {}    variable "instance\_zone" {}    variable "instance\_type" {     default = "e2-micro"   }    variable "instance\_network" {} 13.   To save [variables.tf](http://variables.tf/), click File > Save. 14.   Add the following VM instances to [mynetwork.tf](http://mynetwork.tf/): # Create the mynet-us-vm instance module "mynet-us-vm" {  source           = "./instance"  instance\_name    = "mynet-us-vm"  instance\_zone    = "us-east4-a"  instance\_network = google\_compute\_network.mynetwork.self\_link } # Create the mynet-eu-vm" instance module "mynet-eu-vm" {  source           = "./instance"  instance\_name    = "mynet-eu-vm"  instance\_zone    = "europe-west1-d"  instance\_network = google\_compute\_network.mynetwork.self\_link } 15.   To save [mynetwork.tf](http://mynetwork.tf/), click File > Save. 16.   Verify that [mynetwork.tf](http://mynetwork.tf/) looks like this, including brackets {} Create mynetwork and its resources To rewrite the Terraform configuration files to a canonical format and style. Command:      terraform fmt To initialize Terraform. Command:     terraform init To create an execution plan. Command:      terraform plan To apply the desired changes. Command:      terraform apply To confirm the planned actions, type: yes Task 3. Verify your deployment Verify your network in the Cloud Console In the Cloud Console, on the Navigation menu (), click VPC network > VPC networks. View the mynetwork VPC network with a subnetwork in every region. On the Navigation menu, click VPC network > Firewall. Sort the firewall rules by Network. View the mynetwork-allow-http-ssh-rdp-icmp firewall rule for mynetwork. Verify your VM instances in the Cloud Console On the Navigation menu (), click Compute Engine > VM instances. View the mynet-us-vm and mynet-eu-vm instances. Note the internal IP address for mynet-eu-vm. For mynet-us-vm, click SSH to launch a terminal and connect. To test connectivity to mynet-eu-vm's internal IP address, run the following command in the SSH terminal (replacing mynet-eu-vm's internal IP address with the value noted earlier)             Command: ping -c 3 <Enter mynet-eu-vm's internal IP here>

Please go to this link -> <https://partner.cloudskillsboost.google/> and sign-in to your account. And then please use this link - <https://partner.cloudskillsboost.google/catalog> to search courses and perform the labs. Here you do not require credits to perform the lab.

As you're getting the error so we would like to suggest you some troubleshooting steps which would resolve your issue. Please follow below steps once; \* Use your personal WiFi network and device. \* Clear cookies and caches of your browser. \* Log out from all the accounts before performing the lab. \* Open incognito mode and sign in with those credentials from which you want to perform the lab. \* Do not change accounts while performing the lab. \* Follow the instructions which are provided in the lab. \* Do not use VPN or any proxy servers. \* Use Chrome for the best response. If you still face the same issue then please don't hesitate to contact us back. We'll check again and do the needful.

=========================================================================

ESSENTIAL GOOGLE CLOUD INFRASTRUCTURE: FOUNDATION:

Course Introduction-

[00:02](javascript:;)Hello.

[00:02](javascript:;)I'm Philipp Maier.

[00:05](javascript:;)I'm Mylene Biddle, we're both Course Developers, at Google Cloud and we want to welcome you to Architecting with Compute Engine, a series of three courses.

[00:13](javascript:;)Before we start using all of the different services that Google Cloud Platform, or GCP offers, let's talk about what GCP is.

[00:21](javascript:;)When you look at Google Cloud, you'll see that it's actually part of a much larger ecosystem.

[00:27](javascript:;)This ecosystem consists of open-source software, providers, partners, developers, third-party software, and other Cloud providers.

[00:35](javascript:;)Google is actually a very strong supporter of open-source software.

[00:39](javascript:;)That's right.

[00:39](javascript:;)Now, Google Cloud consists of Chrome, Google devices, Google Maps, Gmail, Google Analytics, G Suite, Google Search, and the Google Cloud Platform.

[00:49](javascript:;)GCP itself is a computing solution platform that really encompasses three core features: infrastructure, platform, and software.

[00:58](javascript:;)This map represents GCP's global infrastructure.

[01:02](javascript:;)As of this recording, GCP's well-provisioned global network connects over 60 zones to over 130 points of presence through a global network of fiber optic cables.

[01:14](javascript:;)And Google is continuously investing in this network, with new regions, points of presence, and subsea cable investments.

[01:21](javascript:;)On top of this infrastructure, GCP uses state of the art software-defined, networking and distributed systems of technologies to host and deliver your services around the world.

[01:31](javascript:;)These technologies are represented by a suite of Cloud-based products and services that is continuously expanding.

[01:38](javascript:;)Now, it's important to understand that there is usually more than one solution for a task or application in GCP.

[01:45](javascript:;)To better understand this, let's look at a solution continuum.

[01:49](javascript:;)Google Cloud Platform spans from infrastructure as a service, or IaaS, to software as a service, or SaaS.

[01:57](javascript:;)You really can build applications on GCP for the web or mobile that are global,

[02:02](javascript:;)auto-scaling, and assistive, and that provide services where the infrastructure is completely invisible to the user.

[02:08](javascript:;)It is not just that Google has opened the infrastructure that powers applications like Search, Gmail, Google Maps, and G Suite.

[02:15](javascript:;)Google has opened all of the services that make these products possible and packaged them for your use.

[02:21](javascript:;)Alternative solutions are possible.

[02:23](javascript:;)For example, you could start up your own VM in Google Compute Engine, install open-source MySQL on

[02:28](javascript:;)it and run it just like a MySQL database on your own computer in a data center.

[02:35](javascript:;)Or you could use the Cloud SQL service, which provides a MySQL instance and handles operational work

[02:41](javascript:;)like backups and security patching for you using the same services Google does to automate backups and patches.

[02:49](javascript:;)You could even move to a NoSQL database that is auto-scaling and serverless so that growth

[02:55](javascript:;)no longer requires adding server instances or possibly changing the design to handle the new capacity.

[03:02](javascript:;)This series of courses focuses on the infrastructure.

[03:05](javascript:;)An IT infrastructure is like a city infrastructure.

[03:09](javascript:;)The infrastructure is the basic underlying framework of fundamental facilities and systems, such as transport, communications, power, water, fuel, and other essential services.

[03:20](javascript:;)The people in the city are like users, and the cars and bikes, and buildings in the city are like applications.

[03:28](javascript:;)Everything that goes into creating and supporting those applications for the users is the infrastructure.

[03:33](javascript:;)The purpose of this course is to explore as efficiently and clearly as possible the infrastructure services provided by GCP.

[03:42](javascript:;)You should become familiar enough with the infrastructure services that you will know what services do and how to use them.

[03:49](javascript:;)We won't go into very deep dive case studies on specific vertical applications.

[03:54](javascript:;)But you'll know enough to put all the building blocks together to build your own solution.

[03:59](javascript:;)Now, GCP offers a range of compute services.

[04:03](javascript:;)The service that might be most familiar to newcomers is Compute Engine, which lets you run virtual machines on-demand in the Cloud.

[04:11](javascript:;)It's Google Cloud's infrastructure as a service solution.

[04:14](javascript:;)It provides maximum flexibility for people who prefer to managed server instances themselves.

[04:20](javascript:;)Google Kubernetes Engine lets you run containerized applications on a cloud environment that Google manages for you under your administrative control.

[04:29](javascript:;)Think of containerization as a way to package code that's designed to be highly portable and to use resources very efficiently.

[04:37](javascript:;)And think of Kubernetes as a way to orchestrate code in containers.

[04:41](javascript:;)App Engine is GCP's fully managed platform as a service framework.

[04:46](javascript:;)That means it's a way to run code in the cloud without having to worry about infrastructure.

[04:52](javascript:;)You just focus on your code and let Google deal with all the provisioning and resource management.

[04:57](javascript:;)You can learn a lot more about App Engine in the "Developing Applications with Google Cloud Platform" course series.

[05:04](javascript:;)Cloud Functions is a completely serverless execution environment or functions as a service.

[05:09](javascript:;)It executes your code in response to events, whether those events occur once a day or many times per second.

[05:16](javascript:;)Google scales resources as required, but you only pay for the service while your code runs.

[05:21](javascript:;)The "Developing Applications with Google Cloud" course series also discusses Cloud Functions.

[05:27](javascript:;)Cloud Run, a managed compute platform that lets you run stateless containers via web requests or Pub/Sub events.

[05:33](javascript:;)Cloud Run is serverless.

[05:34](javascript:;)That means it removes all infrastructure management tasks so you can focus on developing applications.

[05:41](javascript:;)It is built on Knative, an open API and runtime environment built on Kubernetes that gives you freedom to move your workloads across different environments and platforms.

[05:50](javascript:;)It can be fully managed on Google Cloud, on Google Kubernetes Engine, or anywhere Knative runs.

[05:57](javascript:;)Cloud Run is fast.

[05:58](javascript:;)It can automatically scale up and down from zero almost instantaneously, and it charges you only for the

[06:04](javascript:;)resources you use calculated down to the nearest 100 milliseconds, so you‘ll never pay for your over-provisioned resources.

[06:13](javascript:;)In this series of courses, In this series of courses, Compute Engine will be our main focus.

[06:18](javascript:;)The Architecting with Google Compute Engine courses are part of the Cloud Infrastructure learning path.

[06:23](javascript:;)This path is designed for IT professionals who are responsible for implementing, deploying, migrating, and maintaining applications in the cloud.

[06:32](javascript:;)The prerequisite for these courses is the Google Cloud Platform Fundamentals: Core Infrastructure course, which you can find in the link section for this video.

[06:42](javascript:;)The Architecting with Google Compute Engine series consists of three courses.

[06:47](javascript:;)Essential Cloud Infrastructure: Foundation is the first course of the Architecting with Compute Engine series.

[06:54](javascript:;)In that course, we start by introducing you to GCP and how to interact with the GCP Console and Cloud Shell.

[07:01](javascript:;)Next, we'll get into virtual networks and you will create VPC networks and other networking objects.

[07:07](javascript:;)Then we'll take a deep dive into virtual machines, and you will create virtual machines using Compute Engine.

[07:13](javascript:;)Essential Cloud Infrastructure: Core Services is the second course of this series.

[07:18](javascript:;)In that course, we start by talking about Cloud IAM and you will administer Identity and Access Management for resources.

[07:25](javascript:;)Next, we'll cover the different data storage services in GCP, and you will implement some of those services.

[07:31](javascript:;)Then we'll go over resource management, where you will manage and examine billing of GCP resources.

[07:37](javascript:;)Lastly, we'll talk about resource monitoring and you will monitor GCP resources using Stackdriver services.

[07:44](javascript:;)Elastic Cloud Infrastructure: Scaling, and Automation, is the last course of the series.

[07:49](javascript:;)In that course, we start by going over the different options to interconnect networks to enable you to connect your infrastructure to GCP.

[07:57](javascript:;)Next, we'll go over GCP is load balancing and auto-scaling services.

[08:01](javascript:;)Would you will get to explore directly.

[08:04](javascript:;)Then we'll cover infrastructure automation services like Terraform so that you can automate the development of GCP infrastructure services.

[08:12](javascript:;)Lastly, we'll talk about other managed services that you might want to leverage in GCP.

[08:17](javascript:;)Now, our goal for you is to remember and understand the different GCP services and features, and

[08:23](javascript:;)also be able to apply your knowledge, analyze requirements, evaluate different options, and create your own services.

[08:30](javascript:;)That's why these courses include interactive hands-on maps through the Qwiklabs platform.

[08:35](javascript:;)Qwiklabs provisions you with a Google account and credentials, so you can access the GCP console for each lab at no cost.

INTERACTING WITH GOOGLE CLOUD:

Module Overview-

[00:00](javascript:;)In this module, we will provide you with an introduction to GCP by building on what you learned about the GCP infrastructure from the course introduction.

[00:08](javascript:;)This module is focused on how to interact with GCP.

[00:11](javascript:;)In the labs of this module, you will explore both the GCP's graphical user interface and it's command-line interface.

[00:18](javascript:;)You will also deploy a solution from the GCP marketplace without having to manually configure the software, Virtual Machine instances, storage, or network settings.

[00:29](javascript:;)To complete your learning experience, I will provide a quick demo of Projects.

[00:32](javascript:;)Let's get started.

Using Google Cloud-

[00:00](javascript:;)There are four ways you can interact with GCP, and we'll talk about each in turn.

[00:05](javascript:;)There's the Google Cloud Platform console or GCP console, Cloud Shell and the Cloud SDK, the APIs and the Cloud mobile app.

[00:14](javascript:;)The GCP console provides a web-based graphical User Interface that you access through console.cloud.google.com For example, you can view your Virtual Machines and their details as shown on the top.

[00:26](javascript:;)If you prefer to work in a terminal window, the Cloud SDK provides the gcloud command line tool.

[00:32](javascript:;)For example, you can list your Virtual Machines and their details as shown on the bottom with the gcloud compute instances list command.

[00:40](javascript:;)GCP, also provides Cloud Shell which is a browser-based interactive shell environment for GCP that you can access from the GCP console.

[00:48](javascript:;)Cloud Shell is a temporary Virtual Machine with five gigabytes of persistent disk storage that has the Cloud SDK pre-installed.

[00:56](javascript:;)Throughout this course, you will apply what you learn in different labs.

[01:00](javascript:;)These labs will have instructions to use the GCP console such as on the navigation menu click ''Compute Engine > VM instances''.

[01:09](javascript:;)Let me dissect these instructions.

[01:11](javascript:;)First, within the GCP console, you will click on the icon with the three horizontal lines, which is the navigation menu as shown on the left.

[01:20](javascript:;)This opens a menu as shown on the right.

[01:23](javascript:;)All of the major products and services are listed on this menu.

[01:27](javascript:;)Then within the menu, hover over Compute Engine to open a sub-menu.

[01:32](javascript:;)Finally, click on VM instances on the sub-menu.

[01:36](javascript:;)You will get more comfortable with these instructions and the GCP console as you work on labs.

[01:41](javascript:;)Now, labs will also use command line instructions.

[01:45](javascript:;)You will enter these instructions either in Cloud Shell or an SSH terminal by simply copying and pasting them.

[01:51](javascript:;)In some cases, you will have to modify these commands, for example when choosing a globally unique name for a Cloud Storage bucket.

[01:59](javascript:;)In addition to the Cloud SDK, you can also use Client Libraries that enable you to easily create and manage resources.

[02:07](javascript:;)GCP Client Libraries expose APIs for two main purposes.

[02:11](javascript:;)App APIs provide access to services, and they're optimized for supported languages such as Node.

[02:17](javascript:;)js or Python.

[02:19](javascript:;)Admin APIs offer functionality for resource management.

[02:22](javascript:;)For example, you can use Admin APIs if you want to build your own automated tools.

[02:27](javascript:;)The Cloud mobile app is another way to interact with GCP.

[02:31](javascript:;)It allows you to manage GCP services from your Android or iOS device.

[02:36](javascript:;)For example, you can start, stop an SSH into Compute Engine instances, and see logs from each instance.

[02:42](javascript:;)You can also set up customizable graphs showing key metrics such as CPU usage, network usage, requests per seconds, and server errors.

[02:52](javascript:;)The app even offers alerts and incident management, and allows you to get up-to-date billing information for your projects, and get billing alerts for projects that are going over budget.

[03:02](javascript:;)You can download the Cloud mobile app from Google Play or from the App Store.

Lab Intro: Working with the Google Cloud Console and Cloud Shell-

[00:00](javascript:;)Slides are great for explaining concepts, but let's apply what we just talked about.

[00:04](javascript:;)In this first lab, you'll explore the GCP peer interface.

[00:08](javascript:;)That is the entry point of the graphical user interface that's called the GCP console.

[00:13](javascript:;)Within the GCP console, you will create a storage bucket in Cloud Storage which is Google's unified object storage.

[00:19](javascript:;)Then you'll repeat the same task using Cloud shell, which is the command line interface in GCP.

[00:25](javascript:;)I encourage you to develop familiarity with both the GCP console, and Cloud shell, and to become comfortable moving back and forth between them.

Working with the Google Cloud Console and Cloud Shell- Lab Needed

Lab Review: Working with the Google Cloud Console and Cloud Shell-

[00:00](javascript:;)In this lab, you created a Cloud Storage bucket using both the GCP Console and Cloud Shell within GCP.

[00:07](javascript:;)The GCP Console can do things Cloud Shell can't and vice versa.

[00:11](javascript:;)For example, the GCP Console can keep track of the context of your configuration activities.

[00:17](javascript:;)It can use the Cloud API to determine from the current system state, what options are valid, and it can perform repetitive or more leveraged activities on your behalf.

[00:26](javascript:;)Cloud Shell in contrast offers detailed and precise control, and through its commands, a way to script and automate activities.

[00:34](javascript:;)However, don't think of the constant Cloud Shell as alternatives, think about it as one extremely flexible and powerful interface.

[00:42](javascript:;)You can stick around for our lab walkthrough.

[00:44](javascript:;)But remember, that GCP is user interface can change, so your environment might look slightly different.

[00:50](javascript:;)So here we are in the GCP Console and the first thing we're going to do is create a bucket using the GCP Console.

[00:58](javascript:;)So to do that, I'm going to use the navigation menu which is the icon up here in

[01:02](javascript:;)the top left corner, and I'm currently scroll down to Storage which is here, and click on Browser.

[01:13](javascript:;)What we want to do is create a bucket, so I'm going to click the Create bucket.

[01:18](javascript:;)The first thing we need to do is define a name, and now this name needs to be a globally unique name.

[01:24](javascript:;)So you could for example use your Qwiklab's project ID here, so that's what I'll do, copy that and paste it in there.

[01:37](javascript:;)The instructions just say to create, you could also choose a change the default storage class is currently set to multi-regional.

[01:44](javascript:;)We'll talk more about that in a later module.

[01:46](javascript:;)You can control the access to the objects and there are even some advanced settings around encryption.

[01:52](javascript:;)So I'm just going to go ahead and click Create.

[01:56](javascript:;)You can see that this now has created a bucket and here we see the bucket ID or the name.

[02:05](javascript:;)So now we're going to access Cloud Shell.

[02:08](javascript:;)Then what we're going to do this we're going to click this button up here on the right corner,

[02:11](javascript:;)says Activate Cloud Shell, and then it will prompt you to start clutches, so we'll click that as well.

[02:19](javascript:;)You can see that's coming up here.

[02:22](javascript:;)You could actually expand this and open this in a new tab, or you could realigned this to get a little bit more real estate in here.

[02:31](javascript:;)So we created a bucket using the GCP Console, now we're going to repeat the same using Cloud Shell.

[02:37](javascript:;)So I'm going to go ahead and copy the command from the lab instructions and paste it in here.

[02:44](javascript:;)Another command has the bucket name here in brackets and we want to change that.

[02:49](javascript:;)So this again has to be a globally unique name.

[02:53](javascript:;)So what we could do is we could again grab the ID of our project and maybe just add something to it.

[03:03](javascript:;)We could just add -shell to say that this is the one that we created from Cloud Shell.

[03:10](javascript:;)So the command is gsutil, these are the commands for Cloud Storage and mb is the make bucket command.

[03:19](javascript:;)You'll see that it has created that here, and we can see if we navigate in the GCP Console back to buckets, that we now have two buckets in here.

[03:28](javascript:;)So we're able to create both of those.

[03:31](javascript:;)So there are other Cloud Shell features that we can explore here.

[03:34](javascript:;)So while we're in Cloud Shell, we can click these three dots over here and get some more options.

[03:41](javascript:;)One of which is we can upload a file, and if I click that, I'm just

[03:45](javascript:;)present it with my browser, and I could for example, select this text file and click Open.

[03:52](javascript:;)We see that it's being uploaded and now that has finished, and then I can use the ls command to list that file.

[04:00](javascript:;)So here's that file.

[04:02](javascript:;)There's also a read me already in there.

[04:04](javascript:;)Then we could copy that now, that file to the bucket that we have.

[04:09](javascript:;)So there's a command for that also in the lab instructions.

[04:12](javascript:;)So again, we're working with Cloud Storage.

[04:14](javascript:;)So gsutil is going to be the command and CP to copy.

[04:19](javascript:;)We're going to give the name of the file, so MyFile.txt, and then we want to get to that Cloud Storage bucket.

[04:28](javascript:;)So we could choose either of the two buckets we've created.

[04:31](javascript:;)Why don't we choose the one we create a from Cloud Shell, paste it in there, and then it's telling us that it's copying over the files.

[04:42](javascript:;)If we now go into that, we can see that now that file is in there.

[04:48](javascript:;)The file doesn't contain anything, so that's why it is that size.

[04:53](javascript:;)Then we could also go ahead and close Cloud Shell, and do some other activities.

[04:59](javascript:;)Task 5 of the lab goes into creating a persistent state in Cloud Shell.

[05:05](javascript:;)So you could open Cloud Shell and we could list for example, all the variable regions with the G loud command that's listed

[05:13](javascript:;)in there, G Cloud compute regions list, and from these regions we can now select a region and store that in an environment variable.

[05:24](javascript:;)So let's take the command from the lab instructions and for class region equals, and let's say for example I pick the US Central 1 region, could paste

[05:35](javascript:;)that in there, store it, and then I could verify that with the echo command, just running that and it's not telling me that that is stored in there.

[05:49](javascript:;)The other thing we could do is we could expand this a little bit, we could also create a folder in here with the MK

[05:59](javascript:;)direct command, and now we could create a configuration file, and then we can append the environment variable that we just created to to that file.

[06:16](javascript:;)Then we could add another one for example, we could also store our project ID.

[06:22](javascript:;)So I can put that in there, grab my project ID, copy that, and store that in the environment variable, and then I

[06:31](javascript:;)run the command from the lab instructions to also append the value of the project ID to my environment variable and the configuration file.

[06:42](javascript:;)Then I can just verify all of that and make sure that that's been stored.

[06:48](javascript:;)So this gives us a method to create environment variables and easily recreate them as Cloud Shell is cycled.

[06:54](javascript:;)However, you will still need to remember to issue this source command each time Cloud Shell is opened.

[07:00](javascript:;)So let's modify the.profilefile so that the source command is issued automatically anytime a terminal Cloud Shell is opened.

[07:07](javascript:;)So we're going to close and reopen Cloud Shell.

[07:13](javascript:;)So let me do that, close it and then reopen it, and then I'm going to paste the echo command again.

[07:24](javascript:;)We see that it's not outputting anything, so that command is coming out down.

[07:28](javascript:;)So let's modify that.profilefile using nano, and at the end of that file, let's go all the way to the bottom.

[07:41](javascript:;)We'll go into paste in sourceinfraclassconfig, and then we're going to save that file to profile, and then exit.

[07:57](javascript:;)Then let's verify that we are able to get that environment variable, that is project ID.

[08:07](javascript:;)So that's currently not in there, that is because I haven't restarted it, propagates run when I restart, sorry for that.

[08:15](javascript:;)So let me close it, let me reopen it, and then let's verify.

[08:20](javascript:;)There we go.

[08:23](javascript:;)So now we can see that expected value and that's because we edit the d.profilefile.

[08:27](javascript:;)That's it.

[08:30](javascript:;)So we've leveraged in this lab, the GCP Console, we created a Storage bucket, we also created a Storage bucket using Cloud Shell, and then we looked into some

[08:38](javascript:;)features run Cloud Shell in terms of uploading files, than copying those files to the Storage bucket, and even at the end configuring the profile and setting some environment variables.

[08:50](javascript:;)That's the end of the lab.

Lab Intro: Infrastructure Preview-

[00:00](javascript:;)In this lab, you're going to experience the power of GCP Automation by setting up a complete Jenkins Continuous Integration environment using the GCP marketplace.

[00:10](javascript:;)You will then verify that you can manage the service from the Jenkins UI and administer the service from the VM host through SSH.

[00:17](javascript:;)Now, you could accomplish a very similar result through manual configuration in a couple of hours or

[00:23](javascript:;)even days, but in this lab you will see it set up in only a few minutes.

Infrastructure Preview- Lab Needed

Demo: Projects-

[00:00](javascript:;)Let's explore projects which are the key organizer of infrastructure resources and relate these resources to billing accounts.

[00:07](javascript:;)Resources can only be created and consumed within projects in a way that projects isolate related resources from one another.

[00:15](javascript:;)I will demonstrate how to create and delete projects, and switch contexts between projects.

[00:20](javascript:;)Some of these actions cannot be performed in the Qwiklabs environment due to security restrictions.

[00:26](javascript:;)Therefore, I'm going to demonstrate them in my environment.

[00:29](javascript:;)So here I am in the GCP console.

[00:32](javascript:;)You can actually see this is a trial account and you can also create a trial account yourself if you would like to follow along with this.

[00:38](javascript:;)Essentially, what I'm going to do first is go ahead and create a project.

[00:42](javascript:;)So I'm going to click up on my product name up here, and there's this icon up here to create a new project, so let me go click that.

[00:52](javascript:;)Now, the one thing I want to do is, I want to define a project name.

[00:57](javascript:;)So let me just say my new project.

[01:03](javascript:;)You can see that it automatically creates the project ID and project ID is going to be unique versus my name is really not so unique.

[01:12](javascript:;)So let me click "Create" on that.

[01:15](javascript:;)It is now telling me here that is going to create that project.

[01:21](javascript:;)I can follow along with that here in the notification pane.

[01:26](javascript:;)One thing to notice is, when you create a new project, that some of the services that you're going to use may not be initially available.

[01:33](javascript:;)So here, I now have my new project.

[01:36](javascript:;)I could now switch projects.

[01:39](javascript:;)So if I go to my home, for example, I see here the project itself, I could

[01:48](javascript:;)go to the project settings, I could shut that down, or I could switch to a different one.

[01:58](javascript:;)So let me actually change up here to this new project that I created.

[02:03](javascript:;)You go in there and let's follow the process for shutting that down.

[02:08](javascript:;)So I'm going to click on "Shutdown", I wants to make sure that I really want to do that.

[02:12](javascript:;)It's telling me a little bit about what's going to happen when I do this.

[02:15](javascript:;)Specifically, all building in traffic serving will stop, but the shutdown is actually scheduled.

[02:21](javascript:;)So it will take 30 days, and this is in case that you want up maybe undo this.

[02:25](javascript:;)So I need to just retype my project ID, and I can actually copy and paste

[02:29](javascript:;)it in here, and I can click ''Shutdown'', and it should now give me a notification.

[02:38](javascript:;)So here it's telling me when exactly it's going to shut this down, and I can click ''OK'' on that.

[02:46](javascript:;)So now, this is being scheduled for shutdown.

[02:50](javascript:;)So now, I can go back and obviously want to grab in project, it's automatically put me in the sight.

[02:58](javascript:;)Alternatively, if I go home, you'll see that I also have an option up here.

[03:01](javascript:;)It's telling me, hey, you really need to select a project.

[03:04](javascript:;)So lots of different ways to go about.

[03:07](javascript:;)So I could click on that and select a project.

[03:11](javascript:;)Now, I want to show in a second how we can also move switched projects on screen Cloud Shell.

[03:17](javascript:;)So let's actually go ahead and create another project.

[03:21](javascript:;)Let's just call this My Second Project.

[03:26](javascript:;)We can create that as well.

[03:28](javascript:;)They'll start in the background for us.

[03:33](javascript:;)So what I want to do now, as I said I want to go to Cloud Shell.

[03:38](javascript:;)So if I go up here on the right corner, it's Activate Cloud Shell.

[03:40](javascript:;)I'll just click on that.

[03:42](javascript:;)It doesn't ask me if you want to start Cloud Shell because I've already been using Cloud Shell with this user.

[03:49](javascript:;)So it's also telling me that I haven't used my Cloud Shell in awhile, so

[03:53](javascript:;)it has to unarchive my disk and that's going to take a little bit of time.

[03:57](javascript:;)But once that's up, we can actually go use gcloud config list command and we can paste

[04:03](javascript:;)it in, and it's going to give us more information about the configuration that we currently have.

[04:09](javascript:;)That will include the project that we currently have selected.

[04:14](javascript:;)We can actually see the project right here.

[04:18](javascript:;)This is the project I'm working on right now.

[04:20](javascript:;)So if I paste in, I automatically copy that when I clicked on it.

[04:25](javascript:;)So I want to instead type in here gcloud config list.

[04:31](javascript:;)So here, we get some more information.

[04:34](javascript:;)I can also use the grep command in here to directly got my project and there we see this is the project that we're currently using.

[04:45](javascript:;)I could actually now even changed the focus of my GCP console to this new project.

[04:51](javascript:;)You'll see if I run this command again, my focus of Cloud Shell is still focused on this other project that I had.

[05:00](javascript:;)So one thing we could do now, is we could store the project ID maybe in

[05:04](javascript:;)an environment variable and then we could maybe set it so we could swap back and forth.

[05:09](javascript:;)So let me get the project ID, it's right here.

[05:13](javascript:;)I'm going to maybe just store that in an environment variable.

[05:18](javascript:;)Let's just call that my project ID1.

[05:22](javascript:;)So let me grab the project ID, copy that, paste it in there.

[05:29](javascript:;)So now, I have that stored and now I could use the gcloud config set project to define an action to change the project ID.

[05:39](javascript:;)Now, you can see that I have that other project ID listed here.

[05:45](javascript:;)So I can actually see that, and I could also now use it the same gcloud

[05:49](javascript:;)config list command and grep the project, and you'll see that now I'm working with different project.

[05:56](javascript:;)That's how easy it is to create and delete projects, and switch contexts between projects.

Quiz: Interacting with Google Cloud-

What is the difference between the Google Cloud Console and Cloud Shell?

Cloud Shell is a command-line tool, while the Cloud Console is a graphical user interface

The Cloud Console is a graphical user interface and Cloud Shell is a command-line tool. Both tools allow you to interact with Google Cloud. Even though the Cloud Console can do things Cloud Shell can't do and vice-versa, don’t think of them as alternatives, but think of them as one extremely flexible and powerful interface.

Which of the following does not allow you to interact with Google Cloud?

Cloud Explorer

That's correct! There are four ways you can interact with Google Cloud: There’s the Cloud Console, Cloud Shell and the Cloud SDK, the APIs, and the Cloud Mobile App. The Cloud Explorer is not a Google Cloud tool.

Module Review-

[00:00](javascript:;)In this module, we looked at how to use GCP ,which you got to experience firsthand in two short labs.

[00:07](javascript:;)I also gave a demonstration of how to use projects, which are the key organizer of infrastructure resources.

[00:13](javascript:;)Now that you can interact with GCP, it's time to explore two of the foundational components of GCP's infrastructure, virtual networks and virtual machines.

[00:23](javascript:;)So, what are you waiting for?

[00:25](javascript:;)Move on to the next module to learn more.

VIRTUAL NETWORKS:

[00:00](javascript:;)In this module, we will be covering virtual networks.

[00:03](javascript:;)GCP uses a software defined network, that is built on a global fiber infrastructure.

[00:08](javascript:;)This infrastructure makes GCP one of the world's largest and fastest networks.

[00:14](javascript:;)Thinking about resources as services instead of as hardware, will help you understand the options that are available, and their behavior.

[00:22](javascript:;)In this module, we start by introducing Virtual Private Cloud or VPC, which is Google's managed networking functionality, for Euro Cloud Platform resources.

[00:32](javascript:;)Then, we dissect networking into its fundamental components.

[00:36](javascript:;)Which are projects, networks, subnetworks, IP addresses, routes, and firewall rules, along with network pricing.

[00:43](javascript:;)Next, you will explore Google Cloud's network structure in a lab, by creating networks of many different varieties, and exploring the network relationships between them.

[00:53](javascript:;)After that, we will look at common network designs.

[00:56](javascript:;)This map represents Google Cloud.

[00:59](javascript:;)On a high level, Google Cloud consists of regions, which are the icons in blue, points of presence or

[01:05](javascript:;)PoPs, which are the dots in blue, a global private network, which is represented by the blue lines, and services.

[01:13](javascript:;)A region is a specific geographical location where you can run your resources.

[01:18](javascript:;)This map shows several regions that are currently operating, as well as future regions.

[01:23](javascript:;)Regions indicated with blue icons have three zones.

[01:27](javascript:;)Iowa is an exception, where the region called US-Central1 has four zones: US-Central1-A, US-Central1-B, US-Central1-C, and US-Central1-F.

[01:37](javascript:;)For up-to-date information on regions and zones, please refer to the documentation in the slides.

[01:44](javascript:;)The PoPs, are where Google's network is connected to the rest of the internet.

[01:48](javascript:;)Google Cloud can bring its traffic closer to its peers, because it operates an extensive global network of interconnection points.

[01:56](javascript:;)This reduces costs and provides users with a better experience.

[02:00](javascript:;)The network connects regions and PoPs, and is composed of a global network of fiber optic cables with several submarine cable investments.

[02:09](javascript:;)For more information about Google's networking infrastructure, please refer to these slides.

[02:14](javascript:;)Let's start by talking about GCPs network, and specifically Virtual Private Cloud or VPC.

Virtual Private Cloud-

[00:00](javascript:;)With GCP, you can provision your GCP resources, connect them to each other, and isolate them from each other in a Virtual Private Cloud.

[00:07](javascript:;)You can also define fine-grained network and policies within GCP and between GCP and On-premises or other public Clouds.

[00:14](javascript:;)Essentially, VPC is a comprehensive set of Google managed networking objects, which we will explore in detail throughout this module.

[00:22](javascript:;)Let me give you a high-level overview of these objects.

[00:25](javascript:;)Projects are going to encompass every single service that you use including networks.

[00:30](javascript:;)Networks come in three different flavors; default, auto mode, and custom mode.

[00:35](javascript:;)Subnetworks allow you to divide or segregate your environment.

[00:39](javascript:;)Regions in zones represents Google's datacenters and they provide continuous Data Protection and high availability.

[00:46](javascript:;)VPC provides IP addresses for internal and external use along with granular IP address range selections.

[00:54](javascript:;)As for virtual machines, in this module, we will focus on configuring VM instances from a networking perspective.

[01:01](javascript:;)We'll also go over routes and firewall rules.

Projects, Networks and Subnetworks-

[00:00](javascript:;)Let’s start exploring the VPC objects by looking at projects, networks, and subnetworks.

[00:08](javascript:;)Projects are the key organizer of infrastructure resources in Google Cloud.

[00:14](javascript:;)A project associates objects and services with billing.

[00:19](javascript:;)Now, it’s unique that projects actually contain entire networks.

[00:24](javascript:;)The default quota for each project is 15 networks, but you can simply request additional quota using the Google Cloud console.

[00:34](javascript:;)These networks can be shared with other projects, or they can be peered with networks in other

[00:40](javascript:;)projects, both of which we will cover later in the Architecting with Google Compute Engine course series.

[00:47](javascript:;)These networks do not have IP ranges but are simply a construct of all of the individual IP addresses and services within that network.

[00:58](javascript:;)Google Cloud’s networks are global, spanning all available regions across the world that I showed earlier.

[01:04](javascript:;)So, you can have one network that literally exists anywhere in the world—Asia, Europe, Americas—all simultaneously.

[01:14](javascript:;)Inside a network, you can segregate your resources with regional subnetworks.

[01:20](javascript:;)I just mentioned that there are different types of networks: default, auto, and custom.

[01:26](javascript:;)Let’s explore these types of networks in more detail.

[01:30](javascript:;)Every project is provided with a default VPC network with preset subnets and firewall rules.

[01:36](javascript:;)Specifically, a subnet is allocated for each region with non-overlapping CIDR blocks and firewall rules that allow ingress traffic for ICMP,

[01:47](javascript:;)RDP, and SSH traffic from anywhere, as well as ingress traffic from within the default network for all protocols and ports.

[01:57](javascript:;)In an auto mode network, one subnet from each region is automatically created within it.

[02:02](javascript:;)The default network is actually an auto mode network.

[02:06](javascript:;)These automatically created subnets use a set of predefined IP ranges with a /20 mask that can be expanded to /16.

[02:16](javascript:;)All of these subnets fit within the 10.128.0.0/9 CIDR block.

[02:23](javascript:;)Therefore, as new Google Cloud regions become available, new subnets in those regions are automatically added to auto mode networks using an IP range from that block.

[02:34](javascript:;)A custom mode network does not automatically create subnets.

[02:39](javascript:;)This type of network provides you with complete control over its subnets and IP ranges.

[02:44](javascript:;)You decide which subnets to create, in regions you choose, and using IP ranges you specify.

[02:50](javascript:;)These IP ranges cannot overlap between subnets of the same network.

[02:55](javascript:;)Now, you can convert an auto mode network to a custom mode network to take advantage of the control that custom mode networks provide.

[03:04](javascript:;)However, this conversion is one way, meaning that custom mode networks cannot be changed to auto mode networks.

[03:11](javascript:;)So, carefully review the considerations for auto mode networks to help you decide which type of network meets your needs.

[03:19](javascript:;)On this slide, we have an example of a project that contains 5 networks.

[03:23](javascript:;)All of these networks span multiple regions across the world, as you can see on the right.

[03:28](javascript:;)Each network contains separate virtual machines: A, B, C, and D. Because VMs A and B are in the

[03:36](javascript:;)same network, network 1, they can communicate using their internal IP addresses, even though they are in different regions.

[03:44](javascript:;)Essentially, your virtual machines, even if they exist in different locations across the world, take advantage of Google's global fiber network.

[03:52](javascript:;)Those virtual machines appear as though they're sitting in the same rack when it comes to a network configuration protocol.

[03:54](javascript:;)VMs C and D, however, are not in the same network.

[03:56](javascript:;)Therefore, by default, these VMs must communicate using their external IP addresses, even though they are in the same region.

[04:03](javascript:;)The traffic between VMs C and D isn’t actually touching the public internet, but is going through the Google Edge routers.

[04:12](javascript:;)This has different billing and security ramifications that we will explore later.

[04:17](javascript:;)Because VM instances within a VPC network can communicate privately on a global scale, a single

[04:22](javascript:;)VPN can securely connect your on-premises network to your Google Cloud network, as shown in this diagram.

[04:30](javascript:;)Even though the two VM instances are in separate regions (us-west1 and us-east1), they leverage Google’s

[04:36](javascript:;)private network to communicate between each other and to an on-premises network through a VPN gateway.

[04:44](javascript:;)This reduces cost and network management complexity.

[04:46](javascript:;)I mentioned that subnetworks work on a regional scale.

[04:50](javascript:;)Because a region contains several zones, subnetworks can cross zones.

[04:55](javascript:;)This slide has a region, region 1, with two zones, zones A and B. Subnetworks can extend across these zones within the same region, such as, subnet-1.

[05:06](javascript:;)The subnet is simply an IP address range, and you can use IP addresses within that range.

[05:12](javascript:;)Notice that the first and second addresses in the range, .0 and .1, are reserved for the network and the subnet’s gateway, respectively.

[05:21](javascript:;)This makes the first and second available addresses .2 and .3, which are assigned to the VM instances.

[05:29](javascript:;)The other reserved addresses in every subnet are the second-to-last address in the range and the last address, which is reserved as the "broadcast" address.

[05:36](javascript:;)To summarize, every subnet has four reserved IP addresses in its primary IP range.

[05:38](javascript:;)Now, even though the two virtual machines in this example are in different zones, they still communicate with each other using the same subnet IP address.

[05:40](javascript:;)This means that a single firewall rule can be applied to both VMs, even though they are in different zones.

[05:42](javascript:;)I mentioned that subnetworks work on a regional scale.

[05:43](javascript:;)Because a region contains several zones, subnetworks can cross zones.

[05:44](javascript:;)This slide has a region, region 1, with two zones, zones A and B. Subnetworks can extend across these zones within the same region, such as, subnet-1.

[05:46](javascript:;)The subnet is simply an IP address range, and you can use IP addresses within that range.

[05:47](javascript:;)Notice that the first and second addresses in the range, .0 and .1, are reserved for the network and the subnet’s gateway, respectively.

[05:49](javascript:;)This makes the first and second available addresses .2 and .3, which are assigned to the VM instances.

[05:51](javascript:;)The other reserved addresses in every subnet are the second-to-last address in the range and the last address, which is reserved as the "broadcast" address.

[05:53](javascript:;)To summarize, every subnet has four reserved IP addresses in its primary IP range.

[05:54](javascript:;)Now, even though the two virtual machines in this example are in different zones, they still communicate with each other using the same subnet IP address.

[05:56](javascript:;)This means that a single firewall rule can be applied to both VMs, even though they are in different zones.

[05:58](javascript:;)Speaking of IP addresses of a subnet, Google Cloud VPCs let you increase the IP address space of any subnets without any workload shutdown or downtime.

[06:09](javascript:;)This diagram illustrates a network with subnets that have different subnet masks, allowing for more instances in some subnets than others.

[06:17](javascript:;)This gives you flexibility and growth options to meet your needs, but there are some things to

[06:21](javascript:;)remember: The new subnet must not overlap with other subnets in the same VPC network in any region.

[06:30](javascript:;)Each IP range for all subnets in a VPC network must be a unique valid CIDR block.

[06:37](javascript:;)Also, the new subnet IP address ranges are regional internal IP addresses and have to fall within valid IP ranges.

[06:46](javascript:;)\* Subnet ranges cannot match, be narrower, or be broader than a restricted range.

[06:52](javascript:;)\* Subnet ranges cannot span a valid RFC range and a privately used public IP address range.

[06:59](javascript:;)\* Subnet ranges cannot span multiple RFC ranges.

[07:02](javascript:;)The new network range must be larger than the original, which means the prefix length value must be a smaller number.

[07:10](javascript:;)In other words, you cannot undo an expansion.

[07:13](javascript:;)Now, auto mode subnets start with a /20 IP range.

[07:19](javascript:;)They can be expanded to a /16 IP range, but no larger.

[07:24](javascript:;)Alternatively, you can convert the auto mode subnetwork to a custom mode subnetwork to increase the IP range further.

[07:31](javascript:;)Also, avoid creating large subnets.

[07:34](javascript:;)Overly large subnets are more likely to cause CIDR range collisions when using Multiple Network Interfaces

[07:39](javascript:;)and VPC Network Peering, or when configuring a VPN or other connections to an on-premises network.

[07:47](javascript:;)Therefore, do not scale your subnet beyond what you actually need.

Demo: Expand a Subnet-

[00:00](javascript:;)Let me show you how to expand a custom subnet within GCP.

[00:04](javascript:;)I've already created a custom subnet with a slash 29 mask.

[00:08](javascript:;)A slash 29 mask provides you with eight addresses.

[00:11](javascript:;)But of those, four are reserved by GCP, which leaves you with another four for your VM instances.

[00:17](javascript:;)Let's try to create another VM instance in this subnet.

[00:21](javascript:;)So here we are on the GCP console, and I have my four instances, and if I go into the network interface details here, you can see that

[00:31](javascript:;)these are part of a network and I have a subnet here, and if I drilled further into that you can see that I currently have a slash 29.

[00:40](javascript:;)So let's go back and try to create that other instance.

[00:44](javascript:;)Just going to click on Create Instance.

[00:46](javascript:;)I don't need a very large machine, I'm okay with the micro, and let's hit Create.

[00:56](javascript:;)Ideally, we should be getting an error now about the fact that the IP space should have been exhausted, so we're just going to wait for that.

[01:05](javascript:;)You can also follow this along in the notification pane up here and see that it is trying to create that right now.

[01:12](javascript:;)So we're going to wait for that and see if we get an error here in a second.

[01:16](javascript:;)Once we have that, we're going to go ahead and expand the subnet.

[01:20](javascript:;)So here we can see that the instance creation has failed, I can hover over this and

[01:23](javascript:;)it's just telling me that the IP space of that subnet has been exhausted just as expected.

[01:30](javascript:;)We actually have a "Retry" button here as well as a notification pane.

[01:34](javascript:;)We're going to try to use that in a second once we expand the subnet to recreate that instance.

[01:39](javascript:;)Now, what's important is to note that all of these four instances are currently running.

[01:42](javascript:;)So we're not going to take any of these town during the subnet expansion.

[01:46](javascript:;)Now to expand the subnet, I could go to VPC networks through the navigation menu,

[01:51](javascript:;)or I can go back by clicking on nic0 here directly through the network interface details.

[01:57](javascript:;)So the subnet, this is what I want to change, so let me click the "Edit" button, and lets expand this

[02:02](javascript:;)all the way to a slash 23, and this is going to allow a lot of instances, actually over 500 instances.

[02:10](javascript:;)We're going to wait for this to update, and then we're going to head back and try to recreate that instance.

[02:20](javascript:;)So we can also follow this process along right here.

[02:24](javascript:;)It's still saving, so we're going to just hang on tight here.

[02:28](javascript:;)It should just take a couple seconds.

[02:31](javascript:;)All right.

[02:31](javascript:;)We see it's complete.

[02:34](javascript:;)Now, I still have that "Retry" button here to recreate that instance.

[02:38](javascript:;)So let me actually click that, and I can head back to Compute Engine to see if that is going to succeed.

[02:48](javascript:;)So here we are, instance five, it's being staged and will soon begin running.

[02:53](javascript:;)Let's see if this works out.

[02:55](javascript:;)We can see that already has now an internal IP address allocated now that we've expanded

[02:59](javascript:;)the subnet itself, and if I refresh this we can see that the instance is now created.

[03:08](javascript:;)That's how easy it is to expand a subnet in GCP without any workload shutdown or downtime

IP Addresses-

[00:00](javascript:;)Person: Now that we've covered Google Cloud Networks at a high level, let's go deeper by exploring the IP addresses.

[00:06](javascript:;)In Google Cloud, each virtual machine can have two IP addresses assigned.

[00:11](javascript:;)One of them is an internal IP address, which is going to be assigned via DHCP internally.

[00:16](javascript:;)Every VM that starts up and any service that depends on virtual machines gets an internal IP address.

[00:21](javascript:;)An example of such services are App Engine and Google Kubernetes Engine, which are explored in other courses.

[00:28](javascript:;)When you create a VM in Google Cloud, its symbolic name is registered with an internal DNS service that translates the name to an internal IP address.

[00:39](javascript:;)The DNS is scoped to the network, so it can translate web URLs and VM names of

[00:44](javascript:;)hosts in the same network, but it can't translate host names from VMs in a different network.

[00:52](javascript:;)The other IP address is the external IP address, but this one is optional.

[00:55](javascript:;)You can assign an external IP address if your device or machine is externally facing.

[01:02](javascript:;)That external IP address can be assigned from a pool, making it ephemeral, or it can be assigned from a reserved external IP address, making it static.

[01:12](javascript:;)If you reserve static external IP address and do not assign it to a resource, such as a VM instance or

[01:17](javascript:;)a forwarding rule, you are charged at a higher rate than for static and ephemeral external IP addresses that are in use.

[01:27](javascript:;)You can use your own publicly routable IP address prefixes as Google Cloud external IP addresses and advertise them on the Internet.

[01:35](javascript:;)In order to be eligible, you must own and bring a /24 block or larger.

Demo: Internal and External IP-

[00:00](javascript:;)I just mentioned that VMs can have internal and external IP addresses.

[00:04](javascript:;)Let's explore this in the GCP Console.

[00:07](javascript:;)So here I am on the Compute Engine page.

[00:11](javascript:;)What I'm going to do is just create a VM and walk through the process of choosing your internal and external IP address.

[00:17](javascript:;)So let me click Create.

[00:19](javascript:;)I can leave the name.

[00:21](javascript:;)You have obviously a selection of regions and zones you can choose, but I want to focus on the IP addresses.

[00:27](javascript:;)So let me go down to this option, expand management security networking sole tenancy.

[00:33](javascript:;)Let's focus on networking.

[00:35](javascript:;)Here at the network interface, I'm going to click the pencil icon.

[00:40](javascript:;)I could choose between two different networks.

[00:42](javascript:;)So if I had different networks, I could choose between them.

[00:44](javascript:;)That's not the case here.

[00:46](javascript:;)Then I have the primary or internal IP and external IP.

[00:50](javascript:;)So if we look at those options, you can see that I can use an ephemeral address either the one that's created automatically or I could custom select one.

[01:01](javascript:;)So within the range that I have here I could just type IP address.

[01:05](javascript:;)I could also reserve a static internal IP address.

[01:08](javascript:;)This is great if you want to keep that IP address for a longer time and we have similar options with the external IP address.

[01:15](javascript:;)But one of the big differences is that you can also just select none.

[01:18](javascript:;)So as I mentioned your instances don't need to have an external IP address.

[01:25](javascript:;)So let's just leave this as ephemeral.

[01:28](javascript:;)By the way, with the slash 20 here we have a lot of space in this IP range over 4,000 addresses.

[01:35](javascript:;)So we could definitely have that many instances.

[01:37](javascript:;)There are also limits of how many instances you can have per network.

[01:41](javascript:;)As of this recording is actually 15,000.

[01:44](javascript:;)So do keep that in mind you might have a very large IP range but that doesn't mean that you actually can create that many instances.

[01:50](javascript:;)That's a quota.

[01:52](javascript:;)There may also be actual limitations on physical hardware that's even available within a specific region or zone.

[01:59](javascript:;)So let me go ahead and create this instance.

[02:03](javascript:;)We're going to keep an eye on the internal and as well as the external IP address.

[02:10](javascript:;)Once the instance is created.

[02:12](javascript:;)Then we're also going to stop and start the instance to see if any of the IP address has changed.

[02:17](javascript:;)So here we can see the internal IP address.

[02:20](javascript:;)So that is definitely within that space that we just looked at.

[02:24](javascript:;)The external IP address obviously is within Google strange here and we could have reserved that, but this is an fMRL one.

[02:32](javascript:;)So let's actually test this out.

[02:33](javascript:;)I'm going to select the instance.

[02:36](javascript:;)I'm going to stop it.

[02:38](javascript:;)So it's telling me that it doesn't move in 90 seconds that might be forced.

[02:43](javascript:;)So if you had any shutdown scripts in here you want to make sure that they can actually complete within 90 seconds.

[02:51](javascript:;)So let's run through that.

[02:53](javascript:;)Remember this external IP address that we currently have here as well as the internal IP address.

[03:03](javascript:;)So this is going to take it's time now.

[03:05](javascript:;)We can also click Refresh to keep an eye on this.

[03:09](javascript:;)But this will take about 90 seconds, and that's just to give your shutdown script enough time to perform any task to gracefully shut down this instance.

[03:20](javascript:;)So here we are, we can see the instance is stopped, the external IP address is gone.

[03:25](javascript:;)So now we're just going to startup that instance again.

[03:28](javascript:;)It's going to tell us it we're going to be build while it's running, that's fine.

[03:32](javascript:;)You can see that the internal IP address remained the same wildest instance stopped.

[03:38](javascript:;)So that has actually stayed for the time being.

[03:41](javascript:;)Now, while this instance spins up which we can by the way monitor the progress over here, we

[03:47](javascript:;)should see that we should be getting a new external IP address now because that was an ephemeral address.

[03:54](javascript:;)So here we can see the instance has started back up and we can see that the external IP address has changed.

[04:00](javascript:;)This demonstrates that every VM needs an internal IP address but external IP addresses are optional and by default, there are ephemeral.

Mapping IP Addresses-

[00:00](javascript:;)Regardless of whether you use an ephemeral or static IP address, the external address is unknown to the OS of the VM.

[00:06](javascript:;)The external IP address is mapped to the VM's internal address transparently by VPC.

[00:11](javascript:;)I am illustrating this here by running ifconfig within a VM in GCP, which only returns the internal IP address.

[00:19](javascript:;)Let's explore this further by looking at DNS resolution for both internal and external addresses.

[00:27](javascript:;)Let's start with internal addresses.

[00:29](javascript:;)Each instance has a host name that can be resolved to an internal IP address.

[00:33](javascript:;)This hostname is the same as the instance name.

[00:36](javascript:;)There's also an internal fully qualified domain name, or FQDN, for an instance that uses the format shown on the slide.

[00:44](javascript:;)If you delete and recreate an instance, the internal IP address can change.

[00:49](javascript:;)This change can disrupt connections from other Compute Engine resources, which must obtain the new IP address before they can connect again.

[00:57](javascript:;)However, the DNS name always points to a specific instance, no matter what the internal IP address is.

[01:03](javascript:;)Each instance has a metadata server that also acts as a DNS resolver for that instance.

[01:10](javascript:;)The metadata server handles all DNS queries for local network resources and routes all other queries to Google's public DNS servers for public name resolution.

[01:20](javascript:;)I previously mentioned that an instance is not aware of any external IP address assigned to it.

[01:26](javascript:;)Instead, the network stores a lookup table that matches external IP addresses with the internal IP addresses of the relevant instances.

[01:35](javascript:;)Now let's look at external addresses.

[01:37](javascript:;)Instances with external IP addresses can allow connections from hosts outside of the project.

[01:43](javascript:;)Users can do so directly using the external IP address.

[01:47](javascript:;)Public DNS records pointing to instances are not published automatically; however, admins can publish these using existing DNS servers.

[01:56](javascript:;)Domain name servers can be hosted on GCP, using Cloud DNS.

[02:01](javascript:;)This is a managed service that is definitely worth considering, so let's explore them more detail.

[02:06](javascript:;)Cloud DNS is a scalable, reliable, and managed authoritive Domain Name System, or DNS, service running on the same infrastructure as Google.

[02:15](javascript:;)Cloud DNS translates requests for domain names like google.com into IP addresses.

[02:21](javascript:;)Cloud DNS uses Google's global network of Anycast name servers to serve your DNS zones from redundant locations around the world, providing lower latency and high availability for your users.

[02:33](javascript:;)High availability is very important because if you can't look up a domain name, the internet might as well be down.

[02:39](javascript:;)That's why GCP offers a 100% uptime Service Level Agreement, or SLA, for domains configured in Cloud DNS.

[02:47](javascript:;)Cloud DNS lets you create and update millions of DNS records without the burden of managing your own DNS servers and software.

[02:55](javascript:;)Instead, you use a simple user interface, command-line interface or API.

[03:00](javascript:;)Another networking feature of GCP is Alias IP Ranges.

[03:03](javascript:;)Alias IP Ranges let you assign a range of internal IP addresses as an alias to a virtual machine's network interface.

[03:12](javascript:;)This is useful if you have multiple services running on a VM, and you want to assign a different IP address to each service.

[03:19](javascript:;)In essence, you can configure multiple IP addresses, representing containers or applications hosted in a VM, without having to define a separate network interface.

[03:29](javascript:;)You just draw the alias IP range from the local subnet's primary or secondary CIDR ranges.

[03:35](javascript:;)This diagram provides a basic illustration of primary and secondary CIDR ranges and VM Alias IP ranges.

Routes and Firewall Rules-

[00:00](javascript:;)So far you've learned about projects, networks, subnetworks, and IP addresses.

[00:04](javascript:;)Let's use what you learned to understand how GCP routes traffic.

[00:08](javascript:;)By default, every network has routes that let instances in a network send traffic directly to each other, even across subnets.

[00:15](javascript:;)In addition, every network has a default route that directs packets to destinations that are outside the network.

[00:22](javascript:;)Although these routes cover most of your normal routing needs, you can also create special routes that overwrite these routes.

[00:29](javascript:;)Just creating a route does not ensure that your packet will be received by the specified next top.

[00:34](javascript:;)Firewall rules must also allow the packet.

[00:37](javascript:;)The default network has pre-configured firewall rules that allow all instances in the network to talk with each other.

[00:44](javascript:;)Manually created networks do not have such rules, so you must create them, as you will experience in the first lab.

[00:51](javascript:;)Routes match packets by destination IP addresses.

[00:55](javascript:;)However, no traffic will flow without also matching a firewall rule.

[00:59](javascript:;)A route is created when a network is created, enabling traffic delivery from "anywhere".

[01:04](javascript:;)Also, a route is created when a subnet is created.

[01:08](javascript:;)This is what enables VMs on the same network to communicate.

[01:12](javascript:;)This slide shows a simplified routing table, but let's look at this in more detail.

[01:17](javascript:;)Each route in the Routes collection may apply to one or more instances.

[01:22](javascript:;)A route applies to an instance if the network and instance tags match.

[01:27](javascript:;)If the network matches and there are no instance tags specified, the route applies to all instances in that network.

[01:35](javascript:;)Compute Engine then uses the Routes collection to create individual read-only routing tables for each instance.

[01:42](javascript:;)This diagram shows a massively scalable virtual router at the core of each network.

[01:47](javascript:;)Every virtual machine instance in the network is directly connected to this router, and all packets leaving

[01:53](javascript:;)a virtual machine instance are first handled at this layer before they are forwarded to their next hop.

[02:00](javascript:;)The virtual network router selects the next hop for a packet by consulting the routing table for that instance.

[02:08](javascript:;)GCP firewall rules protect you virtual machine instances from unapproved connections, both inbound and outbound, known as ingress and egress, respectively.

[02:18](javascript:;)Essentially, every VPC network functions as a distributed firewall.

[02:23](javascript:;)Although firewall rules are applied to the network as a whole, connections are allowed or denied at the instance level.

[02:30](javascript:;)You can think of the firewall as existing not only between your instances and other networks, but between individual instances within the same network.

[02:39](javascript:;)GCP firewall rules are stateful.

[02:41](javascript:;)This means that if a connection is allowed between a source and a target or a target at a destination, all subsequent traffic in either direction will be allowed.

[02:51](javascript:;)In other words, firewall rules allow bidirectional communication once a session is established.

[02:56](javascript:;)Also, if for some reason, all firewall rules in a network are deleted, there is still

[03:00](javascript:;)an implied "Deny all" ingress rule and an implied "Allow all" egress rule for the network.

[03:08](javascript:;)You can express your desired firewall configuration as a set of firewall rules.

[03:13](javascript:;)Conceptually, a firewall rule is composed of the following parameters: The direction of the rule.

[03:19](javascript:;)Inbound connections are matched against ingress rules only, and outbound connections are matched against egress rules only.

[03:26](javascript:;)The source of the connection for ingress packets, or the destination of the connection for egress packets.

[03:32](javascript:;)The protocol and port of the connection, where any rule can be restricted to apply to specific protocols only or specific combinations of protocols and ports only.

[03:42](javascript:;)The action of the rule, which is to allow or deny packets that match the direction, protocol, port, and source or destination of the rule.

[03:50](javascript:;)The priority of the rule, which governs the order in which rules are evaluated.

[03:54](javascript:;)The first matching rule is applied.

[03:56](javascript:;)The rule assignment.

[03:58](javascript:;)By default, all rules are assigned to all instances, but you can assign certain rules to certain instances only.

[04:04](javascript:;)For more information on firewall rule components, please refer to the links section of this video.

[04:10](javascript:;)Let's look at some GCP firewall use cases for both egress and ingress.

[04:16](javascript:;)Egress firewall rules control outgoing connections originated inside your GCP network.

[04:22](javascript:;)Egress allow rules allow outbound connections that match specific protocol, ports, and IP addresses.

[04:29](javascript:;)Egress deny rules prevent instances from initiating connections that match non-permitted port, protocol, and IP range combinations.

[04:37](javascript:;)For egress firewall rules, destinations to which a rule applies may be specified using IP CIDR ranges.

[04:44](javascript:;)Specifically, you can use the destination ranges to protect from undesired connections initiated by a VM instance towards an external host, as shown on the left.

[04:55](javascript:;)You can also use destination ranges to prevent undesired connections from internal VM instances to specific GCP CIDR ranges.

[05:03](javascript:;)This is illustrated in the middle, where a VM in a specific subnet is shown attempting to connect inappropriately to another VM within the same network.

[05:13](javascript:;)Ingress firewall rules protect against incoming connections to the instance from any source.

[05:18](javascript:;)Ingress allow rules allow specific protocol, ports, and IP ranges to connect in.

[05:24](javascript:;)The firewall prevents instances from receiving connections on non-permitted ports and protocols.

[05:30](javascript:;)Rules can be restricted to only affect particular sources.

[05:34](javascript:;)Source CIDR ranges can be used to protect an instance from undesired connections coming either from external networks or from GCP IP ranges.

[05:43](javascript:;)This diagram illustrates a VM receiving a connection from an external address, and another VM receiving a connection from a VM within the same network.

[05:51](javascript:;)You can control ingress connections from a VM instance by constructing inbound connection conditions using source CIDR ranges, protocols, or ports.

Pricing-

[00:00](javascript:;)Before you apply what you just learned, let's talk about network pricing.

[00:04](javascript:;)It is important that you understand the circumstances in which you are built for GCP's network.

[00:09](javascript:;)This table, is from the Compute Engine documentation and it lists the price of each traffic type.

[00:15](javascript:;)First of all, egress or traffic coming into GCP's network is not charged, unless there is a resource, such as a load balancer that is processing egress traffic.

[00:25](javascript:;)Responses to request account as egress and are charged.

[00:30](javascript:;)The rest of this table, lists egress or traffic leaving a virtual machine.

[00:35](javascript:;)Egress traffic to the same zone, is not charged as long as that egress is through the internal IP address of an instance.

[00:43](javascript:;)Also egress traffic to Google products like YouTube, maps, drive, or traffic to a different GCP service within the same region, is not charged for.

[00:52](javascript:;)However, there is a charge for egress between zones in the same region, egress within a

[00:57](javascript:;)zone, if the traffic is through the external IP address of an instance, and egress between regions.

[01:05](javascript:;)As for the difference in egress traffic to the same zone, Compute Engine cannot determine the zone of a virtual machine through the external IP address.

[01:14](javascript:;)Therefore, this traffic is treated like egress between zones in the same region.

[01:19](javascript:;)Also there are some exceptions and pricing can always change.

[01:23](javascript:;)So refer to the documentation in the links section of these slides.

[01:27](javascript:;)Now, you are charged for static and ephemeral external IP addresses.

[01:32](javascript:;)This table, represents the external IP pricing for us-central1 as of this recording.

[01:38](javascript:;)You can see that if you reserve a static external IP address and do not assign it to a resource, such as a VM

[01:45](javascript:;)instance or a forwarding rule, you are charged at a higher rate and for static and ephemeral external IP addresses that are in use.

[01:54](javascript:;)Also external IP addresses on preemptible VMs, have a lower charge than for standard VM instances.

[02:02](javascript:;)Remember, pricing can always change, so please refer to the documentation link in the slides.

[02:08](javascript:;)Also I recommend using the GCP pricing calculator to estimate the cost of a collection of resources, because each GCP service has its own pricing model.

[02:20](javascript:;)The pricing calculator is a web-based tool, that you use to specify the expected consumption of certain services and resources, and it then provides you with an estimated cost.

[02:31](javascript:;)For example, you can specify a specific instance type, in a specific region along with 100 gigabytes of monthly egress traffic to Americas and EMEA.

[02:42](javascript:;)The pricing calculator then returns the total estimated cost.

[02:47](javascript:;)You can adjust the currency and time frame to meet your needs, and when you

[02:51](javascript:;)finish, you can e-mail the estimate or save it to a specific URL for future reference.

[02:57](javascript:;)To use the pricing calculator today, refer to the link in the slides.

Lab Intro: VPC Networking-

[00:00](javascript:;)Let's apply some of the network features we just discussed in a lab.

[00:03](javascript:;)In this lab, you create an auto mode VPC network with firewall rules and to VM instances.

[00:09](javascript:;)Then you convert the auto mode network to a custom mode network and create other custom mode networks as shown in this network diagram.

[00:17](javascript:;)You also explore the connectivity across networks.

VPC Networking- Lab Needed

Lab Review: VPC Networking-

[00:00](javascript:;)In this lab, you explore the default network, and determine that you cannot create VM instances without a VPC network.

[00:06](javascript:;)So you created a new auto mode VPC network, with subnets, roots, firewall rules, and two VM instances, and tested connectivity for those VM instances.

[00:17](javascript:;)Because auto mode networks aren't recommended for production, you converted the auto mode network to a custom mode network.

[00:24](javascript:;)Next, you create two more custom mode VPC networks with firewall rules and VM instances using the GCP console, and the GCloud command line.

[00:34](javascript:;)Then you test the connectivity across VPC networks, which worked when you pinged external IP addresses, but not when you pinged internal IP addresses.

[00:43](javascript:;)VPC networks are, by default, isolated private networking domains.

[00:48](javascript:;)Therefore, no internal IP address communication is allowed between networks, unless you set up mechanisms such as VPC peering, or a VPN connection.

[00:57](javascript:;)You can stay for a lab walk through.

[00:59](javascript:;)But remember, that GCP's user interface can change, so your environment might look slightly different.

[01:05](javascript:;)All right.

[01:05](javascript:;)So here I am in the GCP console.

[01:08](javascript:;)The first thing I'm going to do is I'm just going to explore the default network.

[01:12](javascript:;)So if I, on the left-hand, side click on the navigation menu, and scroll down to VPC network, we will see that this project has a default network.

[01:23](javascript:;)Every project has a default network.

[01:26](javascript:;)That is unless you have an organizational policy that prevents this default network from being created.

[01:33](javascript:;)But essentially, all the different projects that used through Qwiklabs will always have this.

[01:38](javascript:;)So in here, we can see we have a different subnet in each of the different regions.

[01:44](javascript:;)All of these are private IP addresses.

[01:46](javascript:;)I can also go to the routes, and these are established automatically with the networks.

[01:51](javascript:;)So we can see routes between subnets, as well as to the default route, to the Internet.

[01:57](javascript:;)We can even look at the firewall rules.

[01:59](javascript:;)The default network comes with some preset firewall rules to allow ICMP traffic from anywhere.

[02:05](javascript:;)RDP traffic, as well as SSH.

[02:09](javascript:;)Then also, all protocols imports within the network.

[02:13](javascript:;)So this is the range of the network.

[02:16](javascript:;)So we also allow all traffic from within the network itself.

[02:21](javascript:;)So let's go ahead and let's actually delete these firewall rules.

[02:24](javascript:;)I can just check them all right here and delete them.

[02:28](javascript:;)Let's just assume that we want to get rid of everything that's been created for us, and just create our own network instead.

[02:34](javascript:;)So I'm going to go ahead and delete these.

[02:37](javascript:;)I can look at the status up here.

[02:41](javascript:;)We can see that all four are being deleted.

[02:43](javascript:;)It'll update as each as being deleted.

[02:45](javascript:;)Once that is done which is now, I can head to the network, select the default network, and we're also just going to delete that entire network.

[03:01](javascript:;)Once we delete this network, we should see that there should be no routes without a network because there's no use case for them.

[03:09](javascript:;)So let's just wait for the network to be deleted and then we'll verify that.

[03:15](javascript:;)So we can, again, see the progress bar up here, that's deleting, you can also hit refresh, and this should just take a couple seconds.

[03:31](javascript:;)You can see that as I'm refreshing, some of the subnets are disappearing.

[03:36](javascript:;)It's actually just deleting them all the subnets first, and then it's getting rid of the

[03:40](javascript:;)network as a whole, because the network is really nothing else than just a combination of subnets.

[03:45](javascript:;)So all these subnets have to be deleted.

[03:46](javascript:;)There we go.

[03:47](javascript:;)They're all gone now.

[03:47](javascript:;)Now, it's just the network itself that is remaining.

[03:56](javascript:;)If I go to routes, we should see that all the routes already gone, because without the subnets, there's really no need for the routes.

[04:03](javascript:;)If I go back to the network, we should see that any moment now the network itself also disappears.

[04:10](javascript:;)There we go.

[04:11](javascript:;)All right.

[04:13](javascript:;)So without a VPC network now, we shouldn't be able to create any VM instances, containers, or app engine application.

[04:21](javascript:;)Let's actually verify that.

[04:22](javascript:;)I'm going to go to the navigation menu, go to compute engine, and let's just try to create an instance, just going to click create.

[04:32](javascript:;)I'm going to leave everything as its default.

[04:35](javascript:;)If I go actually under networking, we should see that it's going to complain here.

[04:41](javascript:;)If I click on networking, that actually doesn't have a local network available.

[04:46](javascript:;)But let's just click create and see what happens, and it does indeed give us an error, and point out the fact that this tab has an issue.

[04:53](javascript:;)So we clearly cannot create an instance, because again, these instances live in networks, and without a network, we can't create it.

[05:01](javascript:;)So let's hit cancel, and what we're going to do now is we're going to create our own auto mode network.

[05:06](javascript:;)So I'm going to head back to VPC networks.

[05:11](javascript:;)You can pin, by the way, the services.

[05:15](javascript:;)So I'm just going to pin VPC network, compute engine, because we're going to be going back and forth between these.

[05:20](javascript:;)Then within VPC network, we're just now going to create our own network.

[05:25](javascript:;)I can give it a name.

[05:27](javascript:;)I'm going to use the same name that I have in lab instructions, which is My Network.

[05:30](javascript:;)Now I have the option of creating a custom or an automatic.

[05:34](javascript:;)Let's start off by creating an automatic network.

[05:37](javascript:;)So that's going to preset all the distance subnets for us in all the different regions that are available.

[05:43](javascript:;)You can scroll through those and see those all in here.

[05:45](javascript:;)They have a preset to side arrange.

[05:48](javascript:;)You can expand that side arrange later.

[05:51](javascript:;)But again, as an auto network, you don't define the actual IP address range.

[05:57](javascript:;)There are also firewall rules that are available.

[05:59](javascript:;)What's interesting here is you see that there's a deny-all ingress and allow all ingress firewall rule.

[06:06](javascript:;)So these are here by default, and they're actually.

[06:09](javascript:;)You can't even uncheck them.

[06:11](javascript:;)So these are actually with all networks that you create, and you can see that this has the highest party integer, which really means it's a lowest priority.

[06:20](javascript:;)So by default, all ingress traffic is denied, and all ingress traffic is allowed.

[06:26](javascript:;)Unless we create other firewalls to see differently.

[06:29](javascript:;)So if I check all these boxes, we're now allowing ingress traffic for these IP ranges, and these protocols imports.

[06:37](javascript:;)So let's go ahead and click create, and we're going to wait for that network to be created.

[06:45](javascript:;)Then we're going to look at the IP addresses for two of the different regions, and

[06:50](javascript:;)we're going to create instances in those regions, and verify that it's taking those IP addresses.

[06:58](javascript:;)So here, you can see the subnets already all populated here.

[07:02](javascript:;)I can monitor the progress also up here, but this is really done any second now.

[07:07](javascript:;)I'm actually going to start heading over to compute engine, and to create our instances.

[07:13](javascript:;)So let's click create.

[07:16](javascript:;)I'm going to give it a name mynet-us-vm.

[07:22](javascript:;)This is going to be in your central one, specifically, the Zone C. I don't really need a big machine.

[07:29](javascript:;)We're just doing some testing here.

[07:30](javascript:;)So let me just create a micro that reduces the cost a little bit, and I'm going to now click create.

[07:41](javascript:;)Then we're going to repeat.

[07:44](javascript:;)I can close this panel over here.

[07:46](javascript:;)The same workflow and create an instance in Europe.

[07:49](javascript:;)So I'm going to grab the name from the lab instructions for that.

[07:53](javascript:;)I'm going to select the Europe West One region, specifically the Zone 1C.

[07:59](javascript:;)Again, a micro machine which is just a shared-core, and click create for that as well.

[08:05](javascript:;)We can see the US Central 1C machine is already up.

[08:10](javascript:;)We also see the internal IP address that has been provided.

[08:15](javascript:;)Again, there are some reserved IP addresses.

[08:17](javascript:;)The dot zero is reserved as well as the dot one.

[08:20](javascript:;)So in both of these ranges, the dot two is the first available address.

[08:24](javascript:;)Now, we can verify that these are part of the right subnet, if I click on nic0, I go to the network interface details.

[08:33](javascript:;)Here, we can see it's part of the sub-network.

[08:35](javascript:;)Now the sub-network, in this case, has the same name as the network because this is an auto network.

[08:40](javascript:;)Here, we can see that it's part of this range.

[08:44](javascript:;)So 1012800/20.

[08:47](javascript:;)Let's verify that, and that is correct.

[08:51](javascript:;)We are in there with a dot two, and let's verify that the other should be now a 10132.00/20.

[09:00](javascript:;)So again, click on nic0, go to the sub-network, and we can see that's true.

[09:05](javascript:;)You can also see here that this address is reserved for the gateway.

[09:10](javascript:;)All right.

[09:10](javascript:;)So that way, the dot two was really the first usable address within that range.

[09:16](javascript:;)So now, these are on the same network.

[09:19](javascript:;)So let's verify some connectivity between those.

[09:22](javascript:;)I'm going to grab the internal IP address of mynet-eu-vm, just copy that, and then we're going to SSH too this other instance.

[09:34](javascript:;)So again, these instances are in two separate regions but in the same network.

[09:39](javascript:;)So we should be able to ping these addresses now.

[09:43](javascript:;)So if I ping three times using the internal address, then we can see that this works.

[09:51](javascript:;)This works because we have that allow internal firewall rule that we selected earlier.

[09:56](javascript:;)I can actually repeat the same by using the name of the instance.

[10:03](javascript:;)You can see that it's taking that name.

[10:06](javascript:;)It's actually has, here, the fully qualified domain name, and it's just using the IP address for that.

[10:13](javascript:;)So VPC networks have an internal DNS service that allows you to address instances by that DNS names, instead of their internal IP addresses.

[10:20](javascript:;)That's very useful because, well, this internal IP address could change, right?

[10:24](javascript:;)But the name is not going to change.

[10:25](javascript:;)So it's always good to be aware of that, that you can use the fully qualified domain name to ping those.

[10:31](javascript:;)All right.

[10:31](javascript:;)Now we can try this whole thing the other way round.

[10:35](javascript:;)Let me exit this instance, grab the internal IP address of the instance in the US, and SSH to the instance in Europe.

[10:47](javascript:;)We're also going to ping the internal IP address here.

[10:54](javascript:;)We can see that works.

[10:55](javascript:;)We could even now try to ping the external IP address.

[11:01](javascript:;)So that's 34, in my case, 671818, and that works as well.

[11:09](javascript:;)The reason that I'm able to ping the external is because I have firewall rule that allows ICMP externally.

[11:17](javascript:;)I can verify those again.

[11:18](javascript:;)If I click on the network interface details, here I can see all of the firewall rules, and what filters they have, and what protocols, and ports.

[11:29](javascript:;)So this all works fine, and let's assume that this workflow has worked for us but now we

[11:34](javascript:;)have decided that we want to convert the auto mode network that we have to a custom mode network.

[11:41](javascript:;)So let's go ahead and do that.

[11:42](javascript:;)We're going to go to "VPC networks", and we're going to click on "my network", and then we're going

[11:51](javascript:;)to click on "edit", and we're going to change this subnet creation mode from auto to custom, and hit "save".

[12:04](javascript:;)So now we can navigate back.

[12:07](javascript:;)You can see that this is in progress up here.

[12:10](javascript:;)The mode still says "auto".

[12:12](javascript:;)We could have also flipped that here.

[12:13](javascript:;)Let's wait for that to be refreshed, and now we can see that this subnetwork is now a custom subnet.

[12:24](javascript:;)So let's say that this has worked so far, and now we've realized that we need a couple more networks.

[12:29](javascript:;)There's a network diagram in the lab that has two other networks, as well as some instances and everything.

[12:36](javascript:;)So let's go ahead and create those.

[12:38](javascript:;)So now we're going to go to "create a VPC network".

[12:42](javascript:;)We're going to create the management network, and rather than starting with automatic and converting, we're just going to start with the custom net.

[12:50](javascript:;)For that we have to define each of these subnets.

[12:53](javascript:;)The minimum information we need to provide is a name, the region, so let's select "us-central1", and then the IP address range, and then can click "done".

[13:09](javascript:;)Now I can add, if I wanted to, another subnet.

[13:13](javascript:;)But the other thing that's very interesting about this is, I'm creating this right now through the GCP

[13:16](javascript:;)console but you can also create networks, as well as subnets, from the command line using Cloud Shell.

[13:23](javascript:;)If I click down here on command line, I'm actually provided with the commands to do that.

[13:28](javascript:;)The first one just creates the network itself.

[13:32](javascript:;)You don't have to use the project flag in here.

[13:35](javascript:;)So we could just say G Cloud compute, networks create, the name of the network, and the fact that this subnet is a custom mode.

[13:42](javascript:;)Similarly then, we create these subnets which is "networks subnets create" the name of the subnets, add the subnet itself, the name of the network, the region, and the range.

[13:54](javascript:;)So again, that's the sort of minimal information.

[13:56](javascript:;)Let's just click close and "create".

[13:59](javascript:;)We'll create the other one from the command line.

[14:01](javascript:;)So it's creating that network, and in parallel I can go and now activate Cloud Shell by clicking up here in the right corner.

[14:09](javascript:;)Yes, I want to start using Cloud Shell.

[14:11](javascript:;)I'm just going to make this a little bit bigger, and once this is up, we're going to use those commands that

[14:16](javascript:;)we just saw to create first a network, and this is going to be the privatenet, which is also of the mode custom.

[14:27](javascript:;)Once we have that, we're going to create two subnets within that network.

[14:33](javascript:;)So you can see in the console that the other network was created.

[14:38](javascript:;)Privatenet, is being created right now here, and once that is ready we can add the two subnets to that.

[14:47](javascript:;)So there we go.

[14:47](javascript:;)There is the subnet.

[14:49](javascript:;)It's also telling us this is a new network.

[14:52](javascript:;)You don't have any firewall rules, here are some commands if you want to create some firewall rules.

[14:55](javascript:;)We'll do that in a second.

[14:58](javascript:;)Let's just create these subnets in here.

[15:01](javascript:;)So first we're going to create one in the US, and then we're also going to create one in Europe.

[15:08](javascript:;)If you wanted to speed this up you could actually launch another Cloud Shell session.

[15:12](javascript:;)Now that the network is up, you could create these subnets in parallel.

[15:16](javascript:;)But we're just going to wait for this to complete and then we'll paste that command in there.

[15:21](javascript:;)You can monitor all of this in a console.

[15:24](javascript:;)If we click "refresh," there we see it.

[15:26](javascript:;)It's also completed.

[15:26](javascript:;)It just returns, I've done exactly what you told me.

[15:30](javascript:;)Let's create the other one.

[15:32](javascript:;)It didn't copy the command correctly.

[15:36](javascript:;)There we go.

[15:38](javascript:;)This is now in Europe, specifically Europe-west1.

[15:46](javascript:;)Refresh.

[15:47](javascript:;)You see that's already being created there.

[15:52](javascript:;)So we can definitely display all of those in the GCP console.

[15:56](javascript:;)If you click the button over here in Cloud Shell you can actually open this in a new window.

[16:02](javascript:;)This actually opens it in a new tab, that way you preserve your real estate.

[16:05](javascript:;)You can keep focusing on the console, as well as focusing on Cloud Shell.

[16:09](javascript:;)So let me actually create some real estate by just clearing this, and then paste the command to list all the networks with just G Cloud compute networks list.

[16:20](javascript:;)So we can see them, three networks.

[16:22](javascript:;)They're all Custom.

[16:24](javascript:;)We can dig deeper into this by also listing the subnetworks, and using the "sort-by" command to sort these by network.

[16:32](javascript:;)So now we'll see my network has a lot of subnets because it used to be in auto mode.

[16:38](javascript:;)Then, I mentioned that we want subnet and for permanet.

[16:41](javascript:;)We've two subnets.

[16:44](javascript:;)So now we're going to create some firewall rules.

[16:47](javascript:;)So let's click on "Firewall rules" up here.

[16:51](javascript:;)You can see the ones that are already there.

[16:53](javascript:;)Click create "Firewall rule".

[16:55](javascript:;)We'll repeat the same process we did earlier.

[16:57](javascript:;)We'll first create this using the console, and then we'll repeat the firewalls for a different network using Cloud Shell.

[17:04](javascript:;)So let me give it a name.

[17:06](javascript:;)Let's make sure I select the right network that the firewall rule applies to.

[17:12](javascript:;)Let's just do all instances.

[17:14](javascript:;)For the IP ranges select all addresses.

[17:20](javascript:;)I'm allowing, in this case, ICMP, SSH, and RDP.

[17:26](javascript:;)So let me define ICMP, and then 22 for SSH, and 3389 for RDP, and now

[17:37](javascript:;)down here I can click on "command line", and we can see this as one long command.

[17:43](javascript:;)Again, you don't need to define the project flags as gcloud compute, firewalls create.

[17:48](javascript:;)The name of the rule, the fact it's an ingress party, that is also actually default.

[17:53](javascript:;)We could leave that out.

[17:55](javascript:;)Importantly is the name of the network.

[17:56](javascript:;)Action allows default too, and then the rules as well as source ranges.

[18:01](javascript:;)So let's create that in the console, and we'll grab the command from the lab instructions to do the same for either network.

[18:11](javascript:;)So here you can see.

[18:13](javascript:;)We paste that in, and that should now create the other firewall rule for us.

[18:21](javascript:;)We can monitor the firewall rules in the console, as well as in Cloud Shell.

[18:26](javascript:;)So we'll run a command to list all the firewall rules in a second.

[18:30](javascript:;)So they're all created.

[18:32](javascript:;)If we list them, we can see them all here.

[18:37](javascript:;)If I refresh this we can also see them right here.

[18:42](javascript:;)So now it's time to create some more Instances, and then explore the connectivity.

[18:48](javascript:;)So let's head back to compute engine.

[18:51](javascript:;)I'm going to create instances in these new networks I created.

[18:54](javascript:;)So let me click "create instance".

[18:57](javascript:;)I'm actually going to close Cloud Shell for now.

[19:01](javascript:;)Let me just make it smaller.

[19:04](javascript:;)We're going to provide a name, and "US- Central1-c".

[19:08](javascript:;)Small machine is very fine.

[19:11](javascript:;)Now importantly I need to expand this option down here to select the right network.

[19:16](javascript:;)We've three options right now, and it has actually pre-selected that network.

[19:22](javascript:;)That's because from an order, it's listed up top.

[19:26](javascript:;)That is correct.

[19:28](javascript:;)So let's click "done" and there's again a command.

[19:32](javascript:;)There's a lot of information here that we don't need.

[19:35](javascript:;)You'll see that in a second when we run our command, like the BootDisk.

[19:39](javascript:;)We're selecting a lot of standard options.

[19:41](javascript:;)So let's just hit "create", and let's pull the command from the lab that creates the same in a different network.

[19:52](javascript:;)That's "gcloud compute instance create", the name of the instance, the zone, the machine type, the subnet.

[19:58](javascript:;)That is the bare minimum that we need to provide.

[20:02](javascript:;)So let's run that.

[20:02](javascript:;)You can see the other instance is already created.

[20:06](javascript:;)I can refresh this.

[20:08](javascript:;)See that the other Instances are already coming up too, and once Cloud Shell is updated we can list all the instances.

[20:15](javascript:;)Let's do that here.

[20:17](javascript:;)Can sort them by zone, or we could sort them by network.

[20:21](javascript:;)So we can see in one zone here we have an instance, and then in another zone we have three instances.

[20:27](javascript:;)Keep in mind these Instances are in different networks, and we can display that if we go to columns and check "Network",

[20:33](javascript:;)you can see that these Instances, with exception of the "mynet" these are on the same IVPC network, the others are indifferent.

[20:41](javascript:;)That's going to now go into the connectivity that we're going to explore.

[20:45](javascript:;)We're going to try to again ping IP addresses, both external and internal, and see what works.

[20:51](javascript:;)So let me grab the Management USVM external IP address, and we're going to SSH to the "mynet-us-vm".

[21:03](javascript:;)They are in the same zone, but they're in different networks.

[21:07](javascript:;)So let's see if we can ping the external IP address, and then we'll try the internal.

[21:14](javascript:;)So external works.

[21:15](javascript:;)That's because we set up the firewall rules for that.

[21:21](javascript:;)I can also do the same for privatenet.

[21:24](javascript:;)I can plug that IP address in which is 35.188.20.220 That works as well.

[21:35](javascript:;)So you can ping those, even though they are in different networks.

[21:40](javascript:;)Now from an internal perspective I should only be able to ping mynet-uvm which we actually tried earlier already.

[21:47](javascript:;)So let me just hop on the other ones.

[21:50](javascript:;)I'm going to try 10.130.0.2, and we can see that's not leading to anything.

[22:02](javascript:;)We should be getting a 100 percent packet loss, and then we'll try the same from the other one.

[22:15](javascript:;)So 172.16.0.2 and we can see that again isn't working either.

[22:24](javascript:;)So even though this Instance is in the same zone as these other instances I'm trying to ping, the fact that they are in a

[22:31](javascript:;)different network does not allow me to ping on the internal IP, unless we set up other mechanisms such as VPC peering or a VPN.

[22:39](javascript:;)That's the end of the lab.

Common Network Designs-

[00:00](javascript:;)Let's use what we have learned so far and look at common network designs.

[00:03](javascript:;)Now, common is a fairly relative term, while I could spend all day talking about network designs, I have picked a handful of designs that best relate to this module.

[00:14](javascript:;)Let's start by looking at availability.

[00:16](javascript:;)If your application needs increased availability, you can place two virtual machines into multiple zones, but within the same subnet work as shown on this slide.

[00:26](javascript:;)Using a single sub-network allows you to to create a file a rule against the sub-network, in this case, 10.2.0.0/16.

[00:35](javascript:;)Therefore, by allocating VMs on a single subnet to separate zones, you get improved availability without additional security complexity.

[00:44](javascript:;)A regional managed instance group contains instances from multiple zones across the same region, which provides increased availability.

[00:53](javascript:;)Next, let's look at globalization.

[00:56](javascript:;)In the previous design we placed resources in different zones in a single region, which provides isolation for many types of infrastructure, hardware and software failures.

[01:08](javascript:;)Putting resources in different regions as shown on this slide provides an even higher degree of failure independence.

[01:15](javascript:;)This allows you to design robust systems with resources spread across different failure domains.

[01:20](javascript:;)When using a global load balancer like the HTTP load balancer, you can route traffic to the region that is closest to the user.

[01:28](javascript:;)This can result in better latency for users and lower network traffic costs for your project.

[01:33](javascript:;)We'll explore both managed instance groups and load balancers later in this course series.

[01:39](javascript:;)Now, as a general security best practice, I recommend only assigning internal IP addresses to your VM instances whenever possible.

[01:49](javascript:;)Cloud NAT is Google's managed network address translation service.

[01:53](javascript:;)It lets you provision your application instances without public IP addresses, while also allowing them to access the internet in a controlled and efficient manner.

[02:04](javascript:;)This means your private instances can access the internet for updates, patching, configuration management, and more.

[02:12](javascript:;)In this diagram Cloud NAT enables two private instances to access an update server on the Internet, which is referred to as outbound NAT.

[02:20](javascript:;)However, Cloud NAT does not Implement inbound NAT.

[02:25](javascript:;)In other words, hosts outside your VPC network cannot directly access any of the private instances behind the cloud NAT gateway.

[02:33](javascript:;)This helps you keep your VPC networks isolated and secure.

[02:38](javascript:;)Similarly, you should enable private Google access to allow VM instances that only have internal IP addresses to reach the external IP addresses of Google APIs and services.

[02:50](javascript:;)For example, if your private VM instance needs to access a cloud storage bucket, you need to enable private Google access.

[03:00](javascript:;)You enable private Google access on a subnet by subnet basis.

[03:04](javascript:;)As you can see in this diagram, subnet A has private Google access enabled and subnet B has it disabled.

[03:11](javascript:;)This allows VMA one to access Google APIs and services, even though it has no external IP address.

[03:20](javascript:;)Private Google access has no effect on instances that have external IP addresses, that's why VMs A2 and B2 can access Google APIs and services.

[03:32](javascript:;)The only VM that can't access those APIs and services is VM B1.

[03:37](javascript:;)This VM has no public IP address and it is in a subnet where Google private access is disabled.

Lab Intro: Implement Private Google Access and Cloud NAT-

[00:00](javascript:;)Let's apply what we just covered.

[00:02](javascript:;)In this lab, you implement Private Google Access and Cloud NAT for a VM instance that doesn't have an external IP address.

[00:09](javascript:;)Then you verify access to public IP addresses of Google APIs and services and other connections to the Internet.

Implement Private Google Access and Cloud NAT- Lab Needed

Lab Review: Implement Private Google Access and Cloud NAT-

[00:00](javascript:;)In this lab, you created an instance with no external IP address and access it using Cloud IAP.

[00:07](javascript:;)You then enable Private Google Access and configured a NAT gateway and verified that vm-internal can access Google APIs and services and other public IP addresses.

[00:18](javascript:;)VM instances without external IP addresses are isolated from external networks.

[00:24](javascript:;)Using Cloud NAT, these instances can access the Internet for updates and patches, and in some cases for bootstrapping.

[00:31](javascript:;)As a managed service, Cloud NAT provides high availability without user management and intervention.

[00:37](javascript:;)Let me walk you through the lab.

[00:39](javascript:;)Now, remember that the GCP user interface can change.

[00:42](javascript:;)So your environment might look slightly different.

[00:46](javascript:;)So the first thing I'm going to do is create the VM instance.

[00:49](javascript:;)After that, we are also going to have to create a VPC network and some firewall rules.

[00:54](javascript:;)So let me go to navigation menu, scroll down to VPC networks.

[01:01](javascript:;)We're going to create a network and call it privatenet.

[01:09](javascript:;)So I'm going to name it privatenet, keep this subnet creation mode as custom.

[01:15](javascript:;)We're just going to create one subnet in here.

[01:17](javascript:;)We're going to call it privatenet-us.

[01:20](javascript:;)Let's place this in the us-central1 region, as given to us in the instructions.

[01:28](javascript:;)Here we go, and we even have an IP address range for that.

[01:32](javascript:;)Now, we are going to enable Private Google Access later.

[01:35](javascript:;)So you want to keep that off for now.

[01:37](javascript:;)I turned it on by accident.

[01:39](javascript:;)So you can see the effect of it being off.

[01:42](javascript:;)So let me click "Done" and click "Create".

[01:46](javascript:;)Now, I'm going to wait for this network to be created and once it's up and running, we're going to

[01:50](javascript:;)add a firewall rule because we want to allow SSH to the instance that we're going to put on this network.

[01:59](javascript:;)So I can see the network here.

[02:01](javascript:;)A firewall rule is created for networks, so I had to wait for that to be ready.

[02:05](javascript:;)So let me go to firewall rules, create firewall rule, give it a name.

[02:11](javascript:;)Specify that the network is privatenet.

[02:15](javascript:;)Let's just do all instances and sort by IP ranges.

[02:20](javascript:;)Now, rather than just saying, "Hey, you can SSH this instance from anywhere," we are actually going to give it a very specific range.

[02:28](javascript:;)This is because we're using Cloud IAP.

[02:30](javascript:;)So we're going to use a Cloud IAP tunnel, and because of that, we can limit the site range.

[02:36](javascript:;)Now, this is for an SSH connection.

[02:38](javascript:;)So I want to enable TCP port 22, and then click "Create".

[02:51](javascript:;)While this is creating, I can go ahead and create my Compute Engine instance.

[02:57](javascript:;)So let's go to "Compute Engine", click "Create".

[03:02](javascript:;)We're going to give it the name vm-internal.

[03:06](javascript:;)Now, we need to make sure we choose a region for which we've created a subnet.

[03:11](javascript:;)So us-central1, so us-central1-c.

[03:15](javascript:;)I can keep them machine type as my standard, n1-standard-1, 1virtualCPU, and I'm going to scroll down.

[03:24](javascript:;)The important thing is I need to select the actual VPC networks.

[03:26](javascript:;)Let's go to networking.

[03:28](javascript:;)Networking again, we're going to edit the network interfaces.

[03:33](javascript:;)I want to select the privatenet network.

[03:37](javascript:;)It only has one subnet, and I'm going to set the external IP address to none.

[03:43](javascript:;)Click "Done" and click "Create".

[03:46](javascript:;)So this is a way to create a private instance.

[03:49](javascript:;)Let me close this.

[03:52](javascript:;)That has no external IP address.

[03:55](javascript:;)Now, when the instance comes up, you will see that we won't be able to directly SSH to it because it doesn't have an external IP address.

[04:06](javascript:;)So if we use this, this wouldn't work on us, so instead what we're going to use is, we're going to do an IAP tunnel.

[04:15](javascript:;)For that, we're going to open Cloud Shell.

[04:20](javascript:;)So let me go click "Activate Cloud Shell", and that popped up in a new window.

[04:30](javascript:;)That can certainly happen sometimes.

[04:32](javascript:;)Looks like there's some A, B testing going on here.

[04:37](javascript:;)So here I have Cloud Shell, doesn't look like it has the correct project set.

[04:44](javascript:;)So let's actually do that.

[04:46](javascript:;)I'm going to set the project and then just grab the project ID from here, and set this up for the correct project, and there we can see that now.

[05:03](javascript:;)So it's setup, and now what I'm going to do is, I'm going to run the command to SSH from here.

[05:10](javascript:;)I'm going to specify this is through IAP, and then I want to confirm.

[05:20](javascript:;)For passphrase, we're just going to hit "Enter" and then "Enter" again.

[05:25](javascript:;)Once this is complete, we should now see that the command prompt has changed to vm-internal.

[05:34](javascript:;)So we're now in vm-internal, it doesn't have an external IP address, but let's confirm that we can't just ping the World Wide Web.

[05:42](javascript:;)This ping command isn't working because vm-internal does not have an external IP address.

[05:50](javascript:;)So we can wait for this to complete and it's failing.

[05:54](javascript:;)Again, when instances don't have external IP addresses, they can only be reached by other instances on the network, either through a managed

[06:01](javascript:;)VPN gateway or Cloud IAP tunnel, and Cloud IAP enables contexts where access to VMs through SSH and RDP without a bastion host.

[06:11](javascript:;)That would be the other idea or option.

[06:13](javascript:;)We could create a bastion host, but that would still have an external IAP.

[06:16](javascript:;)Then we're just using the bastion host to then connect to this.

[06:19](javascript:;)Instead, we can just use Cloud Shell and IAP.

[06:23](javascript:;)So this isn't working.

[06:25](javascript:;)So what we're going to now is we're going to look into Private Google Access.

[06:29](javascript:;)So currently, VM instance with no external IP address can use Private Google Access to reach external IP addresses of Google APIs and services.

[06:38](javascript:;)But by default, this is disabled.

[06:40](javascript:;)We saw that earlier, we left it as disabled.

[06:42](javascript:;)So let's test the effect of this being disabled.

[06:45](javascript:;)I'm going to go to the navigation menu, and we're going to create a cloud storage bucket.

[06:50](javascript:;)So let's go to "Storage".

[06:53](javascript:;)I'm going to click "Create Bucket".

[06:56](javascript:;)Now, the most difficult piece is you need to have a unique bucket name.

[07:00](javascript:;)You could do that by grabbing the ID of a project.

[07:05](javascript:;)Click Continue, you can leave this as Multi-region, we can leave everything else by default, and just click Create.

[07:12](javascript:;)The important thing is you're going to have to remember that bucket names, so here's the bucket.

[07:18](javascript:;)So I'm going to do now, is I'm going to go back to Cloud Shell.

[07:22](javascript:;)Importantly, I'm still in my VM Instance here.

[07:27](javascript:;)So I want to change that, so let me exit out of here.

[07:31](javascript:;)So now I'm back in Cloud Shell and then I'm going to run a command to copy

[07:38](javascript:;)an image from a public bucket to my bucket, but I need to specify what my bucket is.

[07:43](javascript:;)So I can take the name of the bucket and add that here to copy this image, so that worked.

[07:56](javascript:;)We can go in here and refresh to verify that we now have an image in here.

[08:02](javascript:;)You can actually click on this image and this just shows you how Private Google Access is implemented pending if it's on or off for a network.

[08:13](javascript:;)We're going to explore that a little bit more.

[08:17](javascript:;)So now what we're going to do is, we're going to now try to copy this image, first from Cloud Shell.

[08:26](javascript:;)Well, Cloud Shell has an external IP address, so that is going to work, run that.

[08:37](javascript:;)I need to actually click Enter.

[08:40](javascript:;)Obviously, I didn't specify my bucket, that is on me.

[08:44](javascript:;)So I need to change my bucket, so typical error that you might see.

[08:51](javascript:;)Let me grab the name of the bucket, placed it in there.

[08:57](javascript:;)Let's try that again, okay, that works.

[08:59](javascript:;)We even use Cloud Shell to move this image anyway, so we're able to access Cloud Storage currently through Cloud Shell.

[09:07](javascript:;)Let's go back to our VM internal [inaudible].

[09:10](javascript:;)So we use the same command use earlier to SSH through a IAP tunnel.

[09:16](javascript:;)Here, I can see the command prompt changed.

[09:20](javascript:;)Now, I'm just going to copy the same command here to copy this image, so I

[09:27](javascript:;)don't have to change the bucket name a couple times, and we're going to run that.

[09:33](javascript:;)We should see that this does not work, because currently VM internal can only send traffic within the VPC networks because again, Private Google Access is disabled.

[09:43](javascript:;)So with two options, we can wait for this to fail and give us an error or we can use Control C to just stop the request.

[09:54](javascript:;)So let's actually just stopped this.

[09:58](javascript:;)What we do now is I'm going to able Private Google Access.

[10:02](javascript:;)So let's go back to the Cloud Console, the Navigation menu and I'm going to navigate to my VPC network, specifically privatenet.

[10:18](javascript:;)Private Google Access is enabled at the subnet level.

[10:22](javascript:;)So I'm going to go directly to the subnet, click the Edit icon, scroll down and able Private Google Access or set it to on, click Save.

[10:33](javascript:;)I'm going to wait for this to update and then I'm going to come back to

[10:39](javascript:;)my instance, my SSH lessons through Cloud Shell and just try to run the command again.

[10:45](javascript:;)So it looks like it's all set, you can also see that here.

[10:49](javascript:;)Going back to my SSH window, run that command again, and now it works.

[10:55](javascript:;)So that's how easy it is to enable Private Google Access.

[11:00](javascript:;)So now in this last task of the lab, we're going to configure a Cloud NAT gateway.

[11:04](javascript:;)Now although our Instance here, VM internal can now access certain Google APIs and services without an external IP address, the instance cannot access the Internet for updates and patches.

[11:15](javascript:;)So for that, we're going to configure our Cloud NAT gateway, but again, we're going

[11:18](javascript:;)to try this behavior first without the NAT gateway and then we're going to enable it.

[11:24](javascript:;)So what we're going to do is I'm going to exit here to just get to my Cloud Shell Instance.

[11:32](javascript:;)There we go, you can see the command prompt changed to Cloud Shell.

[11:35](javascript:;)I'm just going to run sudo apt-get update, and that should obviously work for my Cloud Shell instance because it has an external IP address.

[11:44](javascript:;)So we can see it's getting all these packages and that is working just fine.

[11:51](javascript:;)So now that's complete, we're going to use the SSH command again using the IP tunnel to get to VM internal, there we can see this change.

[12:03](javascript:;)Now we're going to run the same command here.

[12:05](javascript:;)You might say, "Well, hold on."

[12:08](javascript:;)It's actually able to get some of these packages.

[12:10](javascript:;)Yes, that's because we've enabled Private Google Access, so it's able to get those within Google.

[12:17](javascript:;)Once it's trying to get something else here, it's failing.

[12:20](javascript:;)So we can just stop that, this is not going to happen.

[12:24](javascript:;)Now we're going to go ahead and configure Cloud NAT gateway and then try to run that command again.

[12:29](javascript:;)So let's go to the Cloud Console and under the Navigation menu, we're going to go to Network services and Cloud NAT.

[12:39](javascript:;)We're going to go click Get started, just give this a name called nat-config.

[12:46](javascript:;)It's just a name that we have in the lab instructions.

[12:50](javascript:;)You really want to follow these lab instructions because any of our labs that are scored, we'll use names that we're defining in the lab instructions.

[12:58](javascript:;)So important distance to be on privatenet, Region is us-central1.

[13:08](javascript:;)For Cloud Router, we currently don't have one, so we're going to go create one.

[13:12](javascript:;)This is super simple, you just give it a name and click Create.

[13:19](javascript:;)Now, there's also a NAT mapping section and this allows you to choose the subnets to map to

[13:24](javascript:;)the NAT gateway, so you could manually assign static IP addresses that should be used when performing that.

[13:29](javascript:;)But in this case, we're not going to go that and get that fancy, we're just going to click Create.

[13:35](javascript:;)We're going to wait now for the gateway status to change to running.

[13:39](javascript:;)So we can see that the status changed to running, it actually only took a couple seconds.

[13:44](javascript:;)Now, even though this is running, it may actually take up to three minutes for the NAT configuration to propagate all the way to the VM.

[13:52](javascript:;)So you want to wait at least a minute before trying to access the Internet again.

[13:57](javascript:;)What I mean by that is in our SSH session that we currently still have to VM internal, we're going to run the command again.

[14:04](javascript:;)I want to make sure it works this time.

[14:05](javascript:;)So I could actually just try it right now and see if it's ready or not.

[14:09](javascript:;)If I do, you see it's still failing at the step.

[14:12](javascript:;)So let me hit Control C and let's get a couple more minutes and then try to run the command again.

[14:18](javascript:;)So we've waited a couple minutes, let's try to run the command one more time and now we can see that's working.

[14:23](javascript:;)It's getting all the packages and with that we can confirm that Cloud NAT decline a gateway is not working.

[14:28](javascript:;)Now, couple of things to remember, the Cloud NAT gateway implements outbound net, but not inbound net.

[14:33](javascript:;)In other words, what that means is that hosts outside of your VPC network, can only respond to connections initiated by your instances.

[14:42](javascript:;)They cannot initiate their own.

[14:44](javascript:;)So new connections to your instances via the net, so keep that in mind.

[14:49](javascript:;)The other thing is in this lab we used IAP, and IAP uses your existing project roles and permissions when you connect to VM instances.

[14:57](javascript:;)So by default, instance owners, which your instance owner since you created this instance.

[15:02](javascript:;)They're the only ones that have the IAP secure tunnel user role.

[15:06](javascript:;)If you want to allow other users to connect to access using VM, using IP tunneling, you need to grant them those roles.

[15:14](javascript:;)You can actually do that directly through the Navigation menu and go to Cloud IP, and just give people those roles.

[15:21](javascript:;)That's the end of the lab.

Quiz: Virtual Networks-

What are the three types of networks offered in Google Cloud?

Default network, auto network, and custom network.

The three network types offered by Google Cloud are: default, auto and custom. Each project starts with a default network. The auto-type network uses the same subnet IP ranges as the default-type, with a network name other than default. A custom-type allows you to specify the IP ranges of subnets.

What is one benefit of applying firewall rules by tag rather than by address?

When a VM is created with a matching tag, the firewall rules apply irrespective of the IP address it is assigned.

When a VM is created the ephemeral external IP address is assigned from a pool. There is no way to predict which address will be assigned, so there is no way to write a rule that will match that VM's IP address before it is assigned. Tags allow a symbolic assignment that does not depend on order in the IP addresses. It makes for simpler, more general, and easier to maintain, firewall rules.

In Google Cloud, what is the minimum number of IP addresses that a VM instance needs?  
One: Only an internal IP address

That's correct! In Google Cloud, each virtual machine needs to have an internal IP address. The external IP address is optional; therefore, a VM instance only needs one IP address.

Module Review-

[00:00](javascript:;)In this module, I gave you an overview of Google's Virtual Private Cloud.

[00:04](javascript:;)We looked at the different objects within VPC like projects, networks, IP addresses, routes, and firewall rules.

[00:10](javascript:;)I also provided a brief overview of how you're network design choices can affect billing.

[00:16](javascript:;)Then you apply the different concepts that we covered in a thorough lab.

[00:20](javascript:;)Next, we looked at common network designs and you got to implement private Google axis and Cloudnet in a lab.

[00:27](javascript:;)Now, that you have a solid understanding of how GCP has implemented networking, let's move on to learn more about other services.

[00:35](javascript:;)Next up, is Compute Engine which offers scalable, high-performance Virtual Machines.

VIRTUAL MACHINES:

Module Overview-

[00:00](javascript:;)In this module, we cover virtual machine instances, or VMs.

[00:04](javascript:;)VMs are the most common infrastructure component and in GCP they're provided by Compute Engine.

[00:09](javascript:;)A VM is similar but not identical to a hardware computer.

[00:13](javascript:;)VMs consists of a virtual CPU, some amount of memory, disk storage, and an IP address.

[00:20](javascript:;)Compute Engine is GCP's service to create VMs.

[00:23](javascript:;)It is very flexible and offers many options including some that can't exist in physical hardware.

[00:28](javascript:;)For example, a micro VM shares a CPU with other virtual machines, so you can get a VM with less capacity at a lower cost.

[00:37](javascript:;)Another example of a function that can't exist in hardware is that some VMs offer burst capability, meaning that

[00:44](javascript:;)the virtual CPU will run above its rated capacity for a brief period, using the available shared physical CPU.

[00:51](javascript:;)The main VM options are CPU, memory, discs, and networking.

[00:56](javascript:;)Now, this is going to be a very robust module; there's a lot of detail to cover here with how virtual machines work on GCP.

[01:03](javascript:;)First, we'll start with the basics of Compute Engine, followed by a quick little lab to get you more familiar with creating virtual machines.

[01:12](javascript:;)Then, we'll look at the different CPU and memory options that enable you to create different configurations.

[01:18](javascript:;)Next, we'll look at images and the different disk options available with Compute Engine.

[01:24](javascript:;)After that, we will discuss very common Compute Engine actions that you might encounter in your day-to-day job.

[01:31](javascript:;)This will be followed by an in-depth lab that explores many of the features and services covered in this module.

[01:37](javascript:;)Let's get started with an overview of Compute Engine.

Compute Engine-

[00:00](javascript:;)As mentioned in the introduction to the course, there is a spectrum of different options in Google Cloud for compute and processing.

[00:07](javascript:;)We will focus on the traditional virtual machine instances.

[00:11](javascript:;)Now the difference is, Compute Engine gives you the utmost in flexibility: run whatever language you want—it's your virtual machine.

[00:20](javascript:;)This is purely an infrastructure as a service or IaaS model.

[00:24](javascript:;)You have a VM and an operating system, and you can choose how to manage it and how

[00:28](javascript:;)to handle aspects, such as autoscaling, where you’ll configure the rules about adding more virtual machines in specific situations.

[00:36](javascript:;)Autoscaling will be covered later in the course.

[00:41](javascript:;)The primary work case of Compute Engine is any generic workload, especially an enterprise application that was designed to run on a server infrastructure.

[00:51](javascript:;)This makes Compute Engine very portable and easy to run in the cloud.

[00:55](javascript:;)Other services, like Google Kubernetes Engine, which consists of containerized workloads, may not be as easily transferable as what you’re used to from on-premises.

[01:05](javascript:;)So what is Compute Engine?

[01:07](javascript:;)At its heart, it's physical servers that you're used to, running inside the Google Cloud environment, with a number of different configurations.

[01:16](javascript:;)Both predefined and custom machine types allow you to choose how much memory and how much CPU you want.

[01:24](javascript:;)You choose the type of disk you want, whether you want to use persistent disks

[01:27](javascript:;)backed up by standard hard drives or solid-state drives, local SSDs, Cloud Storage, or a mix.

[01:35](javascript:;)You can even configure the networking interfaces and run a combination of Linux and Windows machines.

[01:40](javascript:;)We will discuss these options in more detail later in the module.

[01:45](javascript:;)Several different features will be covered throughout this module, such as machine rightsizing, startup scripts, metadata, availability policies, OS patch management, and pricing and usage discounts.

[01:57](javascript:;)Let’s start by looking at the compute options.

[02:00](javascript:;)Compute Engine provides several different machine types that we’ll discuss later in this module.

[02:05](javascript:;)If those machines don’t meet your needs, you can also customize your own machine.

[02:10](javascript:;)Your choice of CPU will affect your network throughput.

[02:14](javascript:;)Specifically, your network will scale at 2 gigabits per second for each CPU core, except for

[02:20](javascript:;)instances with 2 or 4 CPUs which receive up to 10 gigabits per second of bandwidth.

[02:26](javascript:;)As of this recording, there is a theoretical maximum throughput of 100 gigabits per second for an instance with 224 vCPU, when you choose an N2D machine series.

[02:39](javascript:;)When you're migrating from an on-premises setup, you're used to physical cores, which have hyperthreading.

[02:44](javascript:;)On Compute Engine, each virtual CPU (or vCPU) is implemented as a single hardware hyper-thread on one of the available CPU Platforms.

[02:55](javascript:;)For an up-to-date list of all the available CPU platforms, refer to the links section of this video After you pick your compute options, you want to choose your disk.

[03:05](javascript:;)You have three options: Standard, SSD, or local SSD.

[03:10](javascript:;)So basically, do you want the standard spinning hard disk drives (HDDs), or flash memory solid-state drives (SSDs)?

[03:18](javascript:;)Both of these options provide the same amount of capacity in terms of disk size when choosing a persistent disk.

[03:25](javascript:;)Therefore, the question really is about performance versus cost, because there's a different pricing structure.

[03:31](javascript:;)Basically, SSDs are designed to give you a higher number of IOPS per dollar versus standard disks, which will give you a higher amount of capacity for your dollar.

[03:43](javascript:;)Local SSDs have higher throughput and lower latency than SSD persistent disks, because they are attached to the physical hardware.

[03:52](javascript:;)However, the data that you store on local SSDs persists only until you stop or delete the instance.

[03:59](javascript:;)Typically, a local SSD is used as a swap disk, just like you would do if you want

[04:03](javascript:;)to create a ramdisk, but if you need more capacity, you can store those on a local SSD.

[04:12](javascript:;)Standard and non-local SSD disks can be sized up to 257 TB for each instance.

[04:19](javascript:;)The performance of these disks scales with each GB of space allocated.

[04:23](javascript:;)As for networking, we have already seen networking features applied to Compute Engine in the previous module’s lab.

[04:31](javascript:;)We looked at the different types of networks and created firewall rules using IP addresses and network tags.

[04:38](javascript:;)You’ll also notice that you can do regional HTTPS load balancing and network load balancing.

[04:45](javascript:;)This doesn’t require any pre-warming because a load balancer isn't a hardware device that needs to analyze your traffic.

[04:52](javascript:;)A load balancer is essentially a set of traffic engineering rules that are coming into

[04:57](javascript:;)the Google network, and VPC is applying your rules destined to your IP address subnet range.

[05:04](javascript:;)We’ll learn more about load balancers in a later course of the Architecting with Google Compute Engine series.

Demo: Create a VM-

[00:00](javascript:;)Let me give you a quick walk through of the VM instance creation process, and point out CPU, storage, and network options in the GCP Console.

[00:08](javascript:;)So here I am already on the Compute Engine instance page.

[00:12](javascript:;)You can get to here by going to the navigation menu and then clicking on "Compute Engine".

[00:18](javascript:;)As we use this in the course a lot, you might actually want to pin this sometimes so that you can get to it more easily.

[00:23](javascript:;)Then within there I have gone into VM instances.

[00:27](javascript:;)So I just want to again show you some of the options that are available when creating instance.

[00:32](javascript:;)To get started I'm going to click on "Create".

[00:35](javascript:;)The first thing I want to choose is a name, so you have that right up here.

[00:40](javascript:;)Then maybe more importantly is actually where you want the instance to be located.

[00:45](javascript:;)So you have an option of all the different available regions.

[00:48](javascript:;)It has the name of the regions as well as the closest city as to where that region is located.

[00:54](javascript:;)Then within the regions you have different zones that you can choose from.

[00:58](javascript:;)You also see on the right-hand side that there is a cost associated with the current configuration, net cost is going to change as we change the configuration.

[01:07](javascript:;)So for example, if I instead of creating an instance in US Central one, I create one maybe in Europe West One.

[01:14](javascript:;)You will see that the cost is slightly adjusted.

[01:18](javascript:;)So I can try that a couple different ways by choosing different locations.

[01:22](javascript:;)You should see that the cost changes depending on the region that we choose.

[01:28](javascript:;)Now it goes further if I then choose the machine type.

[01:33](javascript:;)There are different types, we'll go into all of those.

[01:36](javascript:;)But if I go in here, the different types explain to me what they provide.

[01:41](javascript:;)This standard n1 standard one provides one virtual CPU with 3.75 gigabytes of memory.

[01:47](javascript:;)If I change to a machine with more CPU and more memory, we'll see that the cost is adjusted.

[01:53](javascript:;)You can also go into details here.

[01:55](javascript:;)It actually spells it out for you that there's a cost for the CPU and memory, but there's also costs for the persistent disk.

[02:01](javascript:;)We haven't configured that yet, but this is the default value and if we configure that, it's going to ingest a cost.

[02:07](javascript:;)There's also sustained use discount would go into that as well, but essentially all of that is what ultimately gets you to this total monthly cost.

[02:16](javascript:;)It's also broken down in an hourly cost here, and we'll talk more about pricing later within the module.

[02:22](javascript:;)So again I can choose different machine types, maybe we want a larger machine type that's going to be more expensive.

[02:27](javascript:;)Maybe I just need a shared core.

[02:29](javascript:;)So something really, a micro machine or a small machine and that can really drive the cost down a lot.

[02:35](javascript:;)So let me go back to the default.

[02:38](javascript:;)The other thing to think about in terms of your region and zone is not just

[02:41](javascript:;)the cost, but really you want to create your instances that are close to your users.

[02:48](javascript:;)Maybe you want to have them spread out across different regions for habilability.

[02:52](javascript:;)You might be having restrictions for data locality meaning that your data has to be in a specific region.

[02:57](javascript:;)So these are all the different things that you want to consider when choosing the region and zone.

[03:01](javascript:;)Now if I scroll further down, one of the next pix options is the boot disk.

[03:06](javascript:;)So we can see here that currently it has a 10 gigabytes standard persistent disk.

[03:11](javascript:;)I can change that.

[03:12](javascript:;)I can change the image itself, but I can also change the boot disk type.

[03:17](javascript:;)Now the boot disk needs to be a persistent disk.

[03:20](javascript:;)We have the standard persistent disk.

[03:21](javascript:;)Think of an HDD.

[03:24](javascript:;)And we have the SSD.

[03:26](javascript:;)You can see that we can define the size here, and you can see that both of them have the same exact maximum size.

[03:34](javascript:;)So if I make this larger, let's say 1,000, then we're going to see that the cost now is adjusted to that disk size.

[03:43](javascript:;)So it can go back, that's very large.

[03:45](javascript:;)Maybe I'm just okay with 10 gigabytes as the boot disk.

[03:49](javascript:;)You can also add more disks.

[03:51](javascript:;)So if I scroll down and go to management security disk networking, I can go to disks here.

[03:58](javascript:;)So here I can choose the type of encryption for the disk.

[04:02](javascript:;)I have Google managed key, Customer managed key, Customer supplied key.

[04:07](javascript:;)Then I can add more disks.

[04:09](javascript:;)So if I add a new disk here then under type I could also choose a local SSD.

[04:16](javascript:;)Disks come in predefined sizes, depending on how many you choose your performance as you can see down here, is going to get adjusted.

[04:24](javascript:;)There is a limit.

[04:26](javascript:;)So at some point the more disk you choose, you're going to hit a limit into your performance, and same if I choose an SSD disk and change the

[04:35](javascript:;)size here, you'll see that also there's a limit but it's also adjusted if I scale as you are changing the IOPS as well as the sustained throughput limit.

[04:47](javascript:;)Now another important thing is obviously networking.

[04:50](javascript:;)So if I click on here, you want to choose the network interface.

[04:54](javascript:;)We already went into this a little bit in the previous module in terms of near choosing your primary internal IP, choosing if you want an external IP or not.

[05:03](javascript:;)So those are all of the different options that you can get there.

[05:07](javascript:;)Now what's really cool is this is all using the GCP Console, but down the road you

[05:10](javascript:;)might say well, I want to create these instances quickly and I want to use a command line.

[05:15](javascript:;)Well, this user interface gives you the command line options.

[05:20](javascript:;)So it's spelling out exactly all the different options you have chosen, how you would recreate that using GCloud.

[05:26](javascript:;)So this can help you get started using the command line and make you more comfortable using that command line.

[05:31](javascript:;)So let me just go ahead and create this instance.

[05:37](javascript:;)Once we create it we have these different columns that are listed here, there are more columns that we can choose from.

[05:44](javascript:;)For example, when you created it, what the machine type is, what network this is a part

[05:51](javascript:;)off, if you had labels or other things, so lots of different options you can list here.

[05:57](javascript:;)So for example, I can just hear when the machine was created, the type as well as the network it is a part off.

[06:04](javascript:;)That's how easy it is to configure the location, CPU, memory, storage, and network interface for a VM instance using the GCP Console.

[06:13](javascript:;)Let's get back to the slides to go over VM axis and lifecycle.

VM Access and Lifecycle-

[00:00](javascript:;)For accessing a VM, the creator of an instance has full root privileges on that instance.

[00:05](javascript:;)On a Linux instance, the creator has SSH capability and can use the Cloud Console to grant SSH capability to other users.

[00:14](javascript:;)On a Windows instance, the creator can use the Cloud Console to generate a username and password.

[00:20](javascript:;)After that, anyone who knows the username a password can connect to the instance using a Remote Desktop Protocol, or RDP, client.

[00:28](javascript:;)We've listed the required firewall rules for both SSH and RDP here, but you don't need

[00:33](javascript:;)to find these if you're using the default network that we covered in the previous module.

[00:39](javascript:;)The lifecycle of a VM is represented by different statuses.

[00:42](javascript:;)We will cover this lifecycle on a high level, but we recommend returning to this diagram as a reference.

[00:49](javascript:;)When you define all the properties of an instance and click Create, the instance enters the provisioning state.

[00:55](javascript:;)Here the resources such as CPU, memory, and disks are being reserved for the instance, but the instance itself isn’t running yet.

[01:04](javascript:;)Next, the instance moves to the staging state where resources have been acquired and the instance is prepared for launch.

[01:11](javascript:;)Specifically, in this state, Compute Engine is adding IP addresses, booting up the system image, and booting up the system.

[01:19](javascript:;)After the instance starts running, it will go through pre-configured startup scripts and enable SSH or RDP access.

[01:26](javascript:;)Now, you can do several things while your instance is running.

[01:31](javascript:;)For example, you can live migrate your virtual machine to another host in the same zone instead of requiring your instance to be rebooted.

[01:39](javascript:;)This allows Google Cloud to perform maintenance that is integral to keeping the infrastructure protected and reliable, without interrupting any of your VMs.

[01:48](javascript:;)While your instance is running, you can also move your VM to a different zone, take a snapshot of the VM’s persistent disk, export the system image, or reconfigure metadata.

[01:58](javascript:;)We will explore some of these tasks in later labs.

[02:01](javascript:;)Some actions require you to stop your virtual machine; for example, if you want to upgrade your machine by adding more CPU.

[02:08](javascript:;)When the instance enters this state, it will go through pre-configured shutdown scripts and end in the terminated state.

[02:14](javascript:;)From this state, you can choose to either restart the instance, which would bring it back to its provisioning state, or delete it.

[02:22](javascript:;)You also have the option to reset a VM, which is similar to pressing the reset button on your computer.

[02:29](javascript:;)This actions wipes the memory contents of the machine and resets the virtual machine to its initial state.

[02:35](javascript:;)The instance remains in the running state through the reset.

[02:39](javascript:;)There are different ways you can change a VM state from running.

[02:44](javascript:;)Some methods involve the Cloud Console and the gcloud command, while others are performed from the OS, such as for reboot and shutdown.

[02:52](javascript:;)It’s important to know that if you are restarting, rebooting, stopping, or even deleting an instance, the shutdown process will take about 90 sec.

[03:01](javascript:;)For a preemptible VM, if the instance does not stop after 30 seconds, Compute Engine sends an ACPI G3 Mechanical Off signal to the operating system.

[03:12](javascript:;)Remember that when writing shutdown scripts for preemptible VMs.

[03:16](javascript:;)As I mentioned previously, Compute Engine can live migrate your virtual machine to another host due to a maintenance event to prevent your applications from experiencing disruptions.

[03:25](javascript:;)A VM’s availability policy determines how the instance behaves in such an event.

[03:29](javascript:;)The default maintenance behavior for instances is to live migrate, but you can change the behavior to terminate your instance during maintenance events instead.

[03:39](javascript:;)If your VM is terminated due to a crash or other maintenance event, your instance automatically restarts by default, but this can also be changed.

[03:49](javascript:;)These availability policies can be configured both during the instance creation and while an instance is running by configuring the Automatic restart and On host maintenance options.

[04:01](javascript:;)For more information on live migration, refer to the links section of this video.

[04:06](javascript:;)OS updates are a part of managing an infrastructure.

[04:08](javascript:;)Let’s see how we can manage the updates to a fleet of Windows VMs.

[04:13](javascript:;)When you provision a premium image, there is a cost associated with the image.

[04:17](javascript:;)This cost includes both the usage of the OS but also the patch management of the OS.

[04:24](javascript:;)Using Google Cloud, we can easily manage the patching of your OSes.

[04:28](javascript:;)Managing patches effectively is a great way to keep your infrastructure up-to-date and reduce the risk

[04:33](javascript:;)of security vulnerabilities But without the right tools, patching can be a daunting and labor intensive process.

[04:39](javascript:;)Use patch management to apply operating system patches across a set of Compute Engine VM instances.

[04:46](javascript:;)Long-running VMs require periodic system updates to protect against defects and vulnerabilities.

[04:52](javascript:;)The OS patch management service has two main components: Patch compliance reporting, which reports insights on the patch status of your VM instances across Windows and Linux distributions.

[05:05](javascript:;)Along with the insights, you can also view recommendations for your VM instances.

[05:11](javascript:;)And Patch deployment, which automates the operating system and software patch update process.

[05:17](javascript:;)A patch job runs across VM instances and applies patches.

[05:21](javascript:;)There are several tasks that can be performed with patch management.

[05:25](javascript:;)You can: Create patch approvals.

[05:27](javascript:;)You can select what patches to apply to your system from the full set of updates available for the specific operating system.

[05:33](javascript:;)You can set up flexible scheduling so you can choose when to run patches (one-time or recurring schedules).

[05:40](javascript:;)Apply advanced patch configuration settings.

[05:43](javascript:;)You can customize your patches by adding configurations such as pre and post patching scripts.

[05:48](javascript:;)And you can manage these patch jobs or updates from a centralized location.

[05:52](javascript:;)When a VM is terminated, you do not pay for memory and CPU resources.

[05:56](javascript:;)However, you are charged for any attached disks and reserved IP addresses.

[06:01](javascript:;)In the terminated state, you can perform any of the actions listed here, such as changing the machine type, but you cannot change the image of a stopped VM.

[06:09](javascript:;)Also, not all of the actions listed here require you to stop a virtual machine.

[06:14](javascript:;)For example, VM availability policies can be changed while the VM is running, as discussed previously.

Lab Intro: Creating Virtual Machines-

[00:00](javascript:;)Let's take some of the Compute Engine constants we just discussed and apply them in a lab.

[00:05](javascript:;)In this lab, you explore virtual machine instance options by creating several standard VMs and a custom VM.

[00:11](javascript:;)You also connect to those VMs using both SSH for Linux machines and RDP for Windows machines.

Creating Virtual Machines- Lab Needed

Lab Review: Creating Virtual Machines-

[00:00](javascript:;)In this lab, you created several Virtual Machine instances of different types with different characteristics.

[00:06](javascript:;)Specifically, you created a small utility VM for administration purposes, a windows VM, and accustomed Linux VM.

[00:14](javascript:;)You also acts as both the Windows and Linux VM and deleted all your creative VMs.

[00:19](javascript:;)In general, start with a smaller VM when you're prototyping solutions to keep the cost down.

[00:25](javascript:;)When you're ready for production, trade up to larger VMs based on capacity.

[00:29](javascript:;)If you building and redundancy for availability, remember to allocate excess capacity to meet performance requirements.

[00:37](javascript:;)Finally, consider using custom VMs when your applications requirements fit between the features of the standard types.

[00:44](javascript:;)You can stay for a lab walk through but remember that GCP user interface can change, so your environment might look slightly different.

[00:52](javascript:;)So in the GCP console, I'm going to navigate to Compute engine and then VM instances, and in here we're going to click "Create".

[01:05](javascript:;)Now, we can define a name there's a small question mark here and if you hover over it can tell you a little bit more about

[01:11](javascript:;)some of the restrictions you have in regards to creating a name, choosing a name that is, and I'm just going to call this my utilityVM.

[01:20](javascript:;)We're going to go over some of the options that actually went over a little bit in the demo, but we obviously can choose region and zones.

[01:28](javascript:;)So let's change the zone to what the lab is instructing, which is 1-C, and then for the machine type we have a lot of different options to choose from.

[01:37](javascript:;)We can see that the cost changes if I scale up to a machine with four

[01:42](javascript:;)virtual CPUs versus a machine that's just maybe a micro, which is a shared core machine.

[01:50](javascript:;)So the cost can change quite drastically.

[01:52](javascript:;)So let's just leave all the remaining settings and click "Create", and once the machine is up and running, we're going to explore the different VM details that we have.

[02:03](javascript:;)So we're going to go into the VM Instances page, and look at things like the CPU platform, the availability policies and so on.

[02:11](javascript:;)So let me do that, let me click on "Utility VM" because it's now in a running state.

[02:17](javascript:;)I'm going to look for a CPU platform, you can see that right here and if I click "Edit", you'll see that I actually am unable to modify that.

[02:27](javascript:;)So that's because I can't do that while the instances is running.

[02:32](javascript:;)There are other things I could do, I could change the firewall rules, I can add network tags.

[02:36](javascript:;)So certain things are available to change while and instances is running.

[02:41](javascript:;)In some cases, you have to stop the instance to change some of the properties.

[02:46](javascript:;)In other cases, you cannot actually even change it unless you delete it.

[02:51](javascript:;)One of those is for example the network interfaces, if you had multiple network interfaces, you'd have to recreate your instance.

[02:57](javascript:;)The good thing is you could keep your boot disk and just reattach that boot disk later on.

[03:02](javascript:;)Now, I can also go and look at the availability policies, just scroll down to some what the enhanced maintenance is.

[03:12](javascript:;)By default, it's set to migrate the VM instance, and that's recommended but you could set this to terminate the instance.

[03:18](javascript:;)It's also going to automatically restart that instance so you could configure that as well.

[03:24](javascript:;)All right, so this is just a little bit exploring the different options, I'm going to go click "Cancel".

[03:30](javascript:;)What we're going to now is explore some of the VM logs.

[03:34](javascript:;)So I'm looking at the detail page here.

[03:38](javascript:;)We want to get a little bit more information about the monitoring options that are available.

[03:44](javascript:;)We can click "Monitoring" here, and we'll get more information about the CPU.

[03:49](javascript:;)This instances barely runs, we don't have much data yet.

[03:52](javascript:;)We get information about the network bytes and packets, disk I/O.

[03:57](javascript:;)We can also, if we go back to details, look at stackdriver logging.

[04:02](javascript:;)So this is now a different user interface and here we now have individual logs that we can explore.

[04:10](javascript:;)We can view options here, we could expand all of these and dig into all of these different

[04:15](javascript:;)logs that are in here and even within there, expand each of the logs to get more information.

[04:22](javascript:;)So this uses stackdriver logging, we'll cover this feature a little bit but more in a later course in the course

[04:27](javascript:;)series if you're interested to learn more about both the logging piece that we just looked at as well as the monitoring.

[04:34](javascript:;)So let's go to Test 2, we're now going to create a windows virtual machine.

[04:39](javascript:;)So I'm going to go back through the navigation menu Compute engine to VM instances, and I'm not going to create a another instance.

[04:50](javascript:;)So I'm going to define a name, and this is just going to call it Windows VM, and we're going to choose a different region and zone this time.

[05:00](javascript:;)Why don't we put this into Europe-West2, and specifically to zone 2A.

[05:07](javascript:;)Let's pick a larger machine.

[05:09](javascript:;)Let's pick one that has two virtual CPUs and 7.5 gigabytes of memory.

[05:15](javascript:;)We can even go ahead now and changed the boot disk because by default, this would be

[05:20](javascript:;)a Linux machine, so if we want to change this because we want to create a windows machine.

[05:26](javascript:;)Specifically, the lab is instructing me to look for the Windows Server 2016 Datacenter Core image.

[05:35](javascript:;)It's first scroll down, I can see that image right here, can change the boot disk.

[05:41](javascript:;)Maybe I want some higher IOPS, I can choose an SSD, and I can even make this larger and click "Select".

[05:50](javascript:;)All of that again is going to affect obviously the cost.

[05:53](javascript:;)I have the cost of the machine, I have the cost of the disc, but the new thing I have now also is

[05:57](javascript:;)the image, I've chosen zupimages which means there is a cost associated with using that image, but it's built all together for you.

[06:06](javascript:;)So you can see that cost broken up right here.

[06:10](javascript:;)Now, the other thing we're going to do is we're going to allow a specific traffic, HTTP and HTTPS traffic.

[06:18](javascript:;)This just creates a network tag for us and then creates filer roles on the network tag so that we can enable traffic on those ports for the TCP protocol.

[06:29](javascript:;)So let's hit "Create" and create this instance.

[06:34](javascript:;)One thing we'll notice when the instance comes up is that under the connect column [inaudible] now seeing an SSH button which is we would have for a Linux machine.

[06:44](javascript:;)We should now see a RDP, which is for the Remote Desktop Protocol.

[06:49](javascript:;)So that's how you would access a Windows machine.

[06:52](javascript:;)Now, the important thing is there you obviously want to configure your username and password so that only authorized users access that machine.

[07:01](javascript:;)So here you can see the RDP button now.

[07:04](javascript:;)What we're going to do now is we're going to click onto the machine and set the Windows password.

[07:11](javascript:;)You can actually also do this by clicking "Down here" set windows password there as well.

[07:17](javascript:;)So actually, let's just do it that way.

[07:19](javascript:;)So you have a username here.

[07:22](javascript:;)It's taking the username that I have for my lab account.

[07:29](javascript:;)So this is the username right now.

[07:31](javascript:;)So I can set that and then it's going to provide me with a password.

[07:43](javascript:;)So there we go.

[07:43](javascript:;)So I can now copy that password and if I use an RDP connection, I can then get into that.

[07:52](javascript:;)This is a little bit outside of the scope for this lab, but if you want

[07:55](javascript:;)to and have an RDP client, you can actually install one through Chrome, through an extension.

[08:00](javascript:;)You could access that instance that way and then configure it and do anything else you wanted to in this Windows Virtual Machine.

[08:08](javascript:;)So let me go ahead and close that, and I'm going to move on to a Task 3, now which is to create a custom Virtual Machine.

[08:16](javascript:;)So I'm going to go back to Create Instance, and to find a name, let's just call it my custom-VM.

[08:25](javascript:;)I'll follow the lab instructions here for setting the region and zone which is US-West1-B.

[08:34](javascript:;)Now, rather than choosing a specific machine type, I can go in here and just select Custom as the machine type and then define the exact numbers of cores memory.

[08:47](javascript:;)So let's say, my specification I want six virtual CPU, and you can see how the scales by the way, there are only certain options.

[08:55](javascript:;)You can choose it goes all the way to 96.

[08:58](javascript:;)So let me choose six here.

[09:03](javascript:;)It's going to scale that memory automatically for us.

[09:06](javascript:;)It gives us a range now depending on that CPU, there's an option to extend

[09:10](javascript:;)the memory so you could get more than 39 and see all the way to 624.

[09:16](javascript:;)This is a separate option, we'll talk more about this in the slides.

[09:21](javascript:;)So let me choose 32, and rather than scrolling here I could also just type the value in and that's also going to adjust the cost now.

[09:32](javascript:;)Sometimes, it's important to note that your custom machine may be between two machine types are actually already provided.

[09:39](javascript:;)A custom machine is generally going to be slightly more expensive.

[09:43](javascript:;)So if you have a standard machine that's very close to the custom machine, it's definitely something you would want to consider.

[09:50](javascript:;)Once the machine runs more than 24 hours, you'll get right sides recommendations.

[09:54](javascript:;)So It'll tell you if the machine is too small or too large and make recommendations based on that.

[09:58](javascript:;)So let's go ahead and create that.

[10:02](javascript:;)Once it's up and running, we're going to SSH to the machine.

[10:07](javascript:;)We're going to run some commands on that machine, and that's actually going to wrap up the Lab for us.

[10:18](javascript:;)Now, with any new project, you get this column here on the right-hand side to help

[10:21](javascript:;)you get started because we're using Qwiklabs generated projects, they're always going to be new products.

[10:28](javascript:;)So you'll see this throughout the training.

[10:29](javascript:;)You can certainly leveraged this if you want but I'm going to collapse that.

[10:33](javascript:;)So VM is up and running, let me SSH to it.

[10:40](javascript:;)Then we're going to run the free command to see information about any unused and used memory and swap space.

[10:48](javascript:;)So let me type free.

[10:50](javascript:;)So we can see that here and that lines up with the memory selections that we made in the machine.

[10:58](javascript:;)I can also see I get more information or details about the RAM installed.

[11:04](javascript:;)So here we get more information about that as well.

[11:10](javascript:;)I can verify the number of processors.

[11:14](javascript:;)So that should have been six, and yep, and prox is sixth, great.

[11:19](javascript:;)We can see details about the CPU itself.

[11:23](javascript:;)So here we get information about the architecture, the byte order, which model exactly, so you can get all this information about any VM that you create.

[11:33](javascript:;)You can also get more information about this in the documentation depending on which region and zone you choose.

[11:38](javascript:;)You'll have different architectures and different models available to choose from.

[11:45](javascript:;)So that's all we wanted to show you here with this lab.

[11:48](javascript:;)We went ahead and created that Virtual Machine, the Utility VM, we created a Windows VM, and then we created a custom

[11:55](javascript:;)Virtual Machine and verified that whatever custom settings we applied were actually used to create the machine by running commands within that machine.

Compute Options-

[00:00](javascript:;)Now that you have completed the lab, let’s dive deeper into the compute options that are available to you in Google Cloud, by focusing on CPU and memory.

[00:07](javascript:;)You have three options for creating and configuring a VM.

[00:13](javascript:;)You can use the Cloud Console as you did in the previous lab, the Cloud Shell command line, or the RESTful API.

[00:20](javascript:;)If you’d like to automate and process very complex configurations, you might want to programmatically configure these through the RESTful API by defining all the different options for your environment.

[00:31](javascript:;)If you plan on using the command line or RESTful API, I recommend that you first configure the instance through the

[00:36](javascript:;)Cloud Console and then ask Compute Engine for the equivalent REST request or command line, as shown in the demo earlier.

[00:46](javascript:;)This way you avoid any typos and get dropdown lists of all the available CPU and memory options.

[00:52](javascript:;)Speaking of CPU and memory options, let’s look at the different machine types that are currently available.

[00:58](javascript:;)When you create a VM, you select a machine type from a machine family that determines the resources available to that VM.

[01:05](javascript:;)There are several machine families you can choose from and each machine family is further organized into machine series and predefined machine types within each series.

[01:15](javascript:;)A machine family is a curated set of processor and hardware configurations optimized for specific workloads.

[01:23](javascript:;)When you create a VM instance, you choose a predefined or custom machine type from your preferred machine family.

[01:31](javascript:;)Alternatively, you can create custom machine types.

[01:35](javascript:;)These let you specify the number of vCPUs and the amount of memory for your instance.

[01:40](javascript:;)There are four Compute Engine machine families.

[01:44](javascript:;)General-purpose Compute-optimized Memory-optimized, and Accelerator-optimized The general-purpose machine family has the best price-performance with the

[01:55](javascript:;)most flexible vCPU to memory ratios, and provides features that target most standard and cloud-native workloads.

[02:05](javascript:;)The E2 machine series is suited for day-to-day computing at a lower cost, especially where there are also no application dependencies on a specific CPU architecture.

[02:16](javascript:;)E2 VMs provide a variety of compute resources for the lowest price on Compute Engine, especially when paired with committed-use discounts.

[02:25](javascript:;)You simply pick the amount of vCPU and memory that you want, and Google provisions it for you.

[02:33](javascript:;)Standard E2 VMs have between 2 to 32 vCPUs with a ratio of 0.5 GB to 8 GB of memory per vCPU.

[02:44](javascript:;)They are a great choice for web servers, small to medium databases, development and test environments, and many applications that don't have strict performance requirements.

[02:55](javascript:;)They offer a compatible performance baseline with the current N1 VMs for those of you who have been using them.

[03:03](javascript:;)The E2 machine series also contains shared-core machine types that use context-switching to share a physical core between vCPUs for multitasking.

[03:14](javascript:;)Different shared-core machine types sustain different amounts of time on a physical core.

[03:19](javascript:;)In general, shared-core machine types can be more cost-effective for running small, non-resource intensive applications than standard, high-memory, or high-CPU machine types.

[03:31](javascript:;)Shared-core E2 machine types have 0.25 to 1 vCPUs with 0.5 GB to 8 GB of memory.

[03:38](javascript:;)N2 and N2D are the next generation following N1 VMs, offering a significant performance jump.

[03:47](javascript:;)N2 and N2D are the most flexible VM types and provide a balance between price and performance

[03:53](javascript:;)across a wide range of VM shapes, including enterprise applications, medium-to-large databases, and many web and app-serving workloads.

[04:03](javascript:;)Committed use and sustained use discounts are supported.

[04:06](javascript:;)N2 VMs support the latest second generation scalable processor from Intel with up to 128 vCPUs and 0.5 to 8 GB of memory per vCPU.

[04:20](javascript:;)Cascade Lake is the default processor for machine types up to 80 vCPUs.

[04:25](javascript:;)For larger machine types Ice Lake is the default processor for specific regions and zones.

[04:31](javascript:;)N2D are AMD-based general purpose VMs.

[04:36](javascript:;)They leverage the latest EPYC Milan and EPYC Rome processors, and provide up to 224 vCPUs per node.

[04:44](javascript:;)Tau T2D VMs are optimized for cost-effective performance of demanding scale-out workloads.

[04:51](javascript:;)T2D VMs are built on the latest 3rd Gen AMD EPYCTM processors and offer full x86 compatibility.

[05:02](javascript:;)They are suited to scale-out workloads including web servers, containerized microservices, media transcoding, and large-scale Java applications.

[05:10](javascript:;)T2D VMs come in predefined VM shapes, with up to 60 vCPUs per VM and 4 GB of memory per vCPU.

[05:20](javascript:;)If you have containerized workloads, Tau VMs are supported by Google Kubernetes Engine to help optimize price-performance.

[05:28](javascript:;)You can add T2D nodes to your GKE clusters by specifying the T2D machine type in your GKE node pools.

[05:37](javascript:;)The compute-optimized machine family has the highest performance per core on Compute Engine and is optimized for compute-intensive workloads.

[05:45](javascript:;)C2 VMs are the best fit VM type for compute-intensive workloads, including AAA gaming, electronic design automation, and high-performance computing across simulations, genomic analysis, or media transcoding.

[06:00](javascript:;)They might also be applications that have very expensive per core licensing and thus would benefit from higher per core performance.

[06:08](javascript:;)Powered by high-frequency Intel-scalable processors, Cascade Lake, C2 machine types offer up to 3.8 Ghz sustained all-core

[06:14](javascript:;)turbo and provide full transparency into the architecture of the underlying server platforms, enabling advanced performance tuning.

[06:24](javascript:;)The C2 series comes in different machine types ranging from 4 to 60 vCPUs, and offers up to 240 GB of memory.

[06:33](javascript:;)You can also attach up to 3 TB of local storage to these VMs for applications that require higher storage performance.

[06:42](javascript:;)The C2D machine series provides the largest VM sizes and are best-suited for high-performance computing (HPC).

[06:49](javascript:;)The C2D series also has the largest available last-level cache (LLC) per core.

[06:56](javascript:;)The C2D machine series come in different machine types ranging from 2 to 112 vCPUs, and offer 4 GB of memory per vCPU.

[07:08](javascript:;)You can also attach up to 3TB of local storage to these machine types for applications that require higher storage performance.

[07:16](javascript:;)C2D VMs are available on the third generation AMD EPYC Milan platform.

[07:24](javascript:;)The memory-optimized machine family provides the most compute and memory resources of any Compute Engine machine family offering.

[07:33](javascript:;)They are ideal for workloads that require higher memory-to-vCPU ratios than the high-memory machine types in the general-purpose machine family.

[07:43](javascript:;)The M1 machine series has up to 4 TB of memory, while the M2 machine series has up to 12 TB of memory.

[07:51](javascript:;)These machine series are well-suited for large in-memory databases such as SAP HANA, as well as in-memory data analytics workloads.

[08:01](javascript:;)Both the M1 and M2 machine series offer the lowest cost per GB of memory on Compute

[08:05](javascript:;)Engine, making them a great choice for workloads that utilize higher memory configurations with low compute resources requirements.

[08:14](javascript:;)Additionally, they offer up to 30% sustained use discounts and are also eligible for committed use discounts, bringing additional savings of greater than 60% for three-year commitments.

[08:27](javascript:;)The accelerator-optimized machine family is ideal for massively parallelized Compute Unified Device Architecture (CUDA) compute workloads, such as machine learning (ML) and high-performance computing (HPC).

[08:40](javascript:;)This family is the optimal choice for workloads that require GPUs.

[08:44](javascript:;)The A2 series has 12 to 96 vCPUs, and up to 1360 GB of memory.

[08:54](javascript:;)Each A2 machine type has a fixed number (up to 16) of NVIDIA’s Ampere A100 GPUs attached.

[09:01](javascript:;)An A100 GPU provides 40 GB of GPU memory—ideal for large language models, databases, and HPC.

[09:12](javascript:;)Additional information, including the latest specs for currently available VM machine types, can be found in the machine types documentation.

[09:20](javascript:;)If none of the predefined machine types match your needs, you can independently specify the number of vCPUs and the amount of memory for your instance.

[09:31](javascript:;)Custom machine types are ideal for the following scenarios: \* When you have workloads that are not a good fit for the predefined machine types that are available to you.

[09:40](javascript:;)\* Or when you have workloads that require more processing power or more memory, but

[09:44](javascript:;)don't need all of the upgrades that are provided by the next larger predefined machine type.

[09:50](javascript:;)It costs slightly more to use a custom machine type than an equivalent predefined machine type, and there are still some limitations in the

[09:56](javascript:;)amount of memory and vCPUs you can select: \* Only machine types with 1 vCPU or an even number of vCPUs can be created.

[10:07](javascript:;)\* Memory must be between 0.9 GB and 6.5 GB per vCPU (by default).

[10:14](javascript:;)\* The total memory of the instance must be a multiple of 256 MB.

[10:22](javascript:;)By default, a custom machine can have up to 6.5 GB of memory per vCPU.

[10:28](javascript:;)However, this might not be enough memory for your workload.

[10:31](javascript:;)At an additional cost, you can get more memory per vCPU beyond the 6.5 GB limit.

[10:38](javascript:;)This is referred to as extended memory, and you can learn more about this in the link provided in the module PDF located in Course Resources.

[10:47](javascript:;)The first thing you want to consider when choosing a region and zone is the geographical location in which you want to run your resources.

[10:55](javascript:;)This map shows the current and planned Google Cloud regions and number of zones.

[11:00](javascript:;)For up-to-date information on the available regions and zones, see the documentation linked in the module PDF located in Course Resources.

[11:07](javascript:;)Each zone supports a combination of Ivy Bridge, Sandy Bridge, Haswell, Broadwell, and Skylake platforms.

[11:15](javascript:;)When you create an instance in the zone, your instance will use the default processor supported in that zone.

[11:21](javascript:;)For example, if you create an instance in the us-central1-a zone, your instance will use a Sandy Bridge processor.

Compute Pricing-

[00:00](javascript:;)Google Cloud offers a variety of different options to keep the prices low for Compute Engine resources.

[00:06](javascript:;)All vCPUs, GPUs, and GB of memory are charged a minimum of 1 minute.

[00:13](javascript:;)For example, if you run your virtual machine for 30 seconds, you will be billed for 1 minute of usage.

[00:19](javascript:;)After 1 minute, instances are charged in 1-second increments.

[00:24](javascript:;)Compute Engine uses a resource-based pricing model, where each vCPU and each GB of memory

[00:29](javascript:;)on Compute Engine is billed separately rather than as a part of a single machine type.

[00:35](javascript:;)You still create instances using predefined machine types, but your bill reports them as individual vCPUs and memory used.

[00:43](javascript:;)There are several discounts available but the discount types cannot be combined.

[00:48](javascript:;)Resource-based pricing allows Compute Engine to apply sustained use discounts to all of your predefined machine types usage in a region collectively rather than to individual machine types.

[01:01](javascript:;)If your workload is stable and predictable, you can purchase a specific amount of vCPUs and memory for a

[01:07](javascript:;)discount off of normal prices in return for committing to a usage term of 1 year or 3 years.

[01:15](javascript:;)The discount is up to 57% for most machine types or custom machine types.

[01:20](javascript:;)The discount is up to 70% for memory-optimized machine types.

[01:25](javascript:;)Preemptible and Spot VMs are instances that you can create and run at a much lower price than normal instances.

[01:33](javascript:;)For both types of VM, Compute Engine might terminate (or preempt) these instances if it requires to access those resources for other tasks.

[01:43](javascript:;)Both preemptive VMs and Spot VMs are excess Compute Engine capacity so their availability varies with usage.

[01:51](javascript:;)Importantly, preemptible VMs can only run for up to 24 hours at a time, but Spot VMs do not have a maximum runtime.

[02:02](javascript:;)The ability to customize the amount of memory and CPU through custom machine types allows for further pricing customization.

[02:11](javascript:;)Speaking of sizing your machine, Compute Engine provides VM sizing recommendations to help you optimize the resource used of your virtual machine instances.

[02:21](javascript:;)When you create a new instance, recommendations for the new instance will appear 24 hours after the instance has been created.

[02:30](javascript:;)Compute Engine also has Free Usage Limits.

[02:33](javascript:;)Sustained use discounts are automatic discounts that you get for running specific Compute Engine resources (vCPUs, memory, and GPU devices) for a significant portion of the billing month.

[02:45](javascript:;)For example, when you run one of these resources for more than 25% of a month,

[02:48](javascript:;)Compute Engine automatically gives you a discount for every incremental minute you use for that instance.

[02:56](javascript:;)The discount increases with usage, and you can get up to 30% net discount for instances that run the entire month.

[03:03](javascript:;)The tables shown on this slide describes the discount you get at each usage level of a VM instance.

[03:10](javascript:;)To take advantage of the full 30% discount, create your VM instances on the first day of the month, because discounts reset at the beginning of each month.

[03:20](javascript:;)The graph on this slide demonstrates how your effective discount increases with use.

[03:25](javascript:;)For example, if you use a virtual machine for 50% of the month, you can an effective discount of 10%.

[03:31](javascript:;)If you use it for 75% of the month, you get an effective discount of 20%.

[03:36](javascript:;)If you use it for 100% of the month, you get an effective discount of 30%.

[03:42](javascript:;)You can also use the Google Cloud Pricing Calculator to estimate your sustained use discount for any arbitrary workload.

[03:49](javascript:;)Compute Engine calculates sustained use discounts based on vCPU and memory usage across each region and separately for each of the following categories: Predefined machine types, and Custom machine types.

[04:03](javascript:;)Let's go through an example where you have two instances that are in the same region but have different machine types and run at different times of the month.

[04:12](javascript:;)Compute Engine breaks down the number of vCPUs and amount of memory used across all instances that

[04:18](javascript:;)use predefined machine types and combines the resources to qualify for the largest sustained usage discounts possible.

[04:27](javascript:;)As shown on this slide, you run the following two instances in the us-central1 region during a month: For

[04:33](javascript:;)the first half of the month, you run an n1-standard-4 instance with 4 vCPUs and 15 GB of memory.

[04:43](javascript:;)For the second half of the month, you run a larger n1-standard-16 instance with 16vCPUs and 60 GB of memory.

[04:55](javascript:;)In this scenario, Compute Engine reorganizes these machine types into individual vCPUs and memory resources and combines their usage to

[05:03](javascript:;)create the following resources, as shown on the bottom: 4 vCPUs and 15 GB of memory for a full month.

[05:13](javascript:;)And then 12 vCPUs and 45 GB of memory for half of the month.

Special Compute Configurations-

[00:00](javascript:;)As I mentioned earlier, a preemptible VM is an instance that you can create and run at a much lower cost than normal instances.

[00:08](javascript:;)See whether you can make your application function completely on preemptible VMs, because a 60 to 91% discount is a significant investment in your application.

[00:18](javascript:;)Now, just to reiterate, these VMs might be preempted at any time, and there is no charge if that happens within the first minute.

[00:27](javascript:;)Also, preemptible VMs are only going to live for up to 24 hours, and you only get a 30-second notification before the machine is preempted.

[00:36](javascript:;)It's also worth noting that there are no live migrations nor automatic restarts in preemptible VMs, but something that we might want to

[00:44](javascript:;)highlight is that you can actually create monitoring and load balancers that can start up new preemptible VMs in case of a failure.

[00:52](javascript:;)In other words, there are external ways to keep restarting preemptible VMs if you need to.

[00:57](javascript:;)One major use case for preemptible VMs is running batch processing jobs.

[01:03](javascript:;)If some of those instances terminate during processing, the job slows but it does not completely stop.

[01:09](javascript:;)Therefore, preemptible instances complete your batch processing tasks without placing additional workload on your existing instances, and without requiring you to pay full price for additional normal instances.

[01:21](javascript:;)Spot VMs are the latest version of preemptible VMs.

[01:26](javascript:;)Spot VMs are virtual machine (VM) instances with the spot provisioning model.

[01:30](javascript:;)New and existing preemptible VMs continue to be supported, and preemptible VMs use the same pricing model as Spot VMs.

[01:39](javascript:;)However, spot VMs provide new features that preemptible VMs do not support.

[01:44](javascript:;)For example, preemptible VMs can only run for up to 24 hours at a time, but Spot VMs do not have a maximum runtime.

[01:53](javascript:;)Like preemptible VMs, Compute Engine might preempt Spot VMs if it needs to reclaim those resources for other tasks.

[02:02](javascript:;)The probability that Compute Engine stops Spot VMs for a system event is generally low, but

[02:06](javascript:;)might vary from a day to day and from zone to zone depending on current conditions.

[02:13](javascript:;)Spot VMs are finite Compute Engine resources, so they might not always be available.

[02:19](javascript:;)Like preemptible VMs, it's worth noting that Spot VMs can't live-migrate to become standard VMs

[02:25](javascript:;)while they are running or be set to automatically restart when there is a maintenance event.

[02:32](javascript:;)There are many best practices which can help you get the most of using Spot VMs.

[02:37](javascript:;)For example, resources for Spot VMs come out of excess and backup Google Cloud capacity.

[02:42](javascript:;)Capacity for spot VMs is often easier to get for smaller machine types, meaning machine types with less resources like vCPU and memory.

[02:53](javascript:;)If you have workloads that require physical isolation from other workloads or virtual machines in order to meet compliance requirements, you want to consider sole-tenant nodes.

[03:03](javascript:;)A sole-tenant node is a physical Compute Engine server that is dedicated to hosting VM instances only for your specific project.

[03:12](javascript:;)Use sole-tenant nodes to keep your instances physically separated from instances in other projects, or to group your instances together on

[03:19](javascript:;)the same host hardware, for example if you have a payment processing workload that needs to be isolated to meet compliance requirements.

[03:29](javascript:;)The diagram on the left shows a normal host with multiple VM instances from multiple customers.

[03:35](javascript:;)A sole-tenant node is shown on the right and it also has multiple VM instances, but they all belong to the same project.

[03:42](javascript:;)You can also fill the node with multiple smaller VM instances of varying sizes, including custom machine types and instances with extended memory.

[03:52](javascript:;)Also, if you have existing operating system licenses, you can bring them to Compute Engine using sole-tenant nodes while minimizing the physical core usage with the in-place restart feature.

[04:04](javascript:;)To learn how to create nodes and place your instances on those nodes, see the link section of this video.

[04:10](javascript:;)Another compute option is to create a shielded VM.

[04:14](javascript:;)Shielded VMs offer verifiable integrity to your VM instances, so you can be confident that your instances haven't been compromised by boot or kernel-level malware or rootkits.

[04:25](javascript:;)Shielded VMS is the first offering in the Shielded Cloud Initiative.

[04:29](javascript:;)The Shielded Cloud Initiative is meant to provide an even more secure foundation for all of Google

[04:33](javascript:;)Cloud by providing verifiable integrity and offering features, like vTPM shielding or sealing, that help prevent data exfiltration.

[04:44](javascript:;)In order to use the shielded VM features, you need to select a shielded image.

[04:49](javascript:;)We'll learn more about images in the next section.

[04:52](javascript:;)Confidential VMs are a breakthrough technology that allows you to encrypt data in use, while it's been processed.

[05:00](javascript:;)Google Cloud's approach to encrypt data in use is simple, easy-to-use deployment without making any code changes to their applications or having to compromise performance.

[05:10](javascript:;)You can collaborate with anyone, all while preserving the confidentiality of your data.

[05:15](javascript:;)Confidential Virtual Machine (Confidential VM) is a type of N2D Compute Engine VM instance running on hosts based on the second generation of AMD Epyc processors, code-named "Rome".

[05:27](javascript:;)Using AMD Secure Encrypted Virtualization (SEV), Confidential VM features built-in optimization of both performance and security for enterprise-class

[05:36](javascript:;)high memory workloads, as well as inline memory encryption that doesn't introduce significant performance penalties on those workloads.

[05:46](javascript:;)The AMD Rome processor family is specifically optimized for compute-heavy workloads, with high memory capacity, high throughput, and support for parallel workloads.

[05:56](javascript:;)In addition, AMD SEV provides for Confidential Computing support.

[06:02](javascript:;)With the confidential execution environments provided by Confidential VM and AMD SEV, Google Cloud keeps customers' sensitive code and other data encrypted in memory during processing.

[06:14](javascript:;)Google does not have access to the encryption keys.

[06:17](javascript:;)You can select the Confidential VM service when creating a new VM using the Google Cloud Console, the Compute Engine API, or the gcloud command-line tool.

Images-

[00:00](javascript:;)Next, let’s focus on images.

[00:02](javascript:;)When creating a virtual machine, you can choose the boot disk image.

[00:06](javascript:;)This image includes the boot loader, the operating system, the file system structure, any pre-configured software, and any other customizations.

[00:15](javascript:;)You can select either a public or custom image.

[00:19](javascript:;)As you saw in the previous lab, you can choose from both Linux and Windows images.

[00:24](javascript:;)Some of these images are premium images, as indicated in parentheses with a p.

[00:29](javascript:;)These images will have per-second charges after a 1-minute minimum, with the exception of SQL Server images, which are charged per minute after a 10-minute minimum.

[00:39](javascript:;)Premium image prices vary with the machine type.

[00:42](javascript:;)However, these prices are global and do not vary by region or zone.

[00:47](javascript:;)You can also use custom images.

[00:49](javascript:;)For example, you can create and use a custom image by pre-installing software that's been authorized for your particular organization.

[00:56](javascript:;)You also have the option of importing images from your own premises or workstation, or from another cloud provider.

[01:03](javascript:;)This is a no-cost service that is as simple as installing an agent, and I highly recommend that you look at it.

[01:09](javascript:;)You can also share custom images with anybody in your project or among other projects, too.

[01:10](javascript:;)A machine image is a Compute Engine resource that stores all the configuration, metadata, permissions, and data from one or more disks required to create a virtual machine (VM) instance.

[01:20](javascript:;)You can use a machine image in many system maintenance scenarios, such as creation, backup and recovery, and instance cloning.

[01:29](javascript:;)Machine images are the most ideal resources for disk backups as well as instance cloning and replication.

Disk Options-

[00:00](javascript:;)At this point you've chosen an operating system, but that operating system is going to be included as part of some kind of disk.

[00:07](javascript:;)So let’s look at the disk options.

[00:09](javascript:;)Every single VM comes with a single root persistent disk, because you're choosing a base image to have that loaded on.

[00:17](javascript:;)This image is bootable in that you can attach it to a VM and boot from it, and it is durable in that it can survive if the VM terminates.

[00:26](javascript:;)To have a boot disk survive a VM deletion, you need to disable the “Delete boot disk when instance is deleted” option in the instance’s properties.

[00:34](javascript:;)As I discussed earlier, there are different types of disks.

[00:39](javascript:;)Let’s explore these in more detail.

[00:42](javascript:;)The first disk that we create is what we call a persistent disk.

[00:47](javascript:;)That means it's going to be attached to the VM through the network interface.

[00:51](javascript:;)Even though it's persistent, it's not physically attached to the machine.

[00:56](javascript:;)This separation of disk and compute allows the disk to survive if the VM terminates.

[01:02](javascript:;)You can also perform snapshots of these disks, which are incremental backups that we’ll discuss later.

[01:08](javascript:;)The choice between HDD and SSD disks comes down to cost and performance.

[01:15](javascript:;)To learn more about disk performance and how it scales with disk size, refer to the link in the course resources for this module.

[01:24](javascript:;)Another cool feature of persistent disks is that you can dynamically resize them, even while they are running and attached to a VM.

[01:31](javascript:;)You can also attach a disk in read-only mode to multiple VMs.

[01:36](javascript:;)This allows you to share static data between multiple instances, which is cheaper than replicating your data to unique disks for individual instances.

[01:47](javascript:;)Zonal persistent disks offer efficient, reliable block storage.

[01:52](javascript:;)Regional persistent disks provide active-active disk replication across two zones in the same region.

[01:59](javascript:;)Regional persistent disks deliver durable storage that is synchronously replicated across zones and are a great option for high-performance databases and enterprise applications that also require high availability.

[02:15](javascript:;)When you configure a zonal or regional persistent disk, you can select one of the following disk types.

[02:21](javascript:;)Standard persistent disks are backed by standard hard disk drives and are suitable for large data processing workloads that primarily use sequential I/Os.

[02:32](javascript:;)Performance SSD persistent disks are backed by solid-state drives and are suitable for enterprise applications and high-performance databases that require lower latency and more IOPS than standard persistent disks provide.

[02:48](javascript:;)Balanced persistent disks are also backed by solid-state drives.

[02:52](javascript:;)They are an alternative to SSD persistent disks that balance performance and cost.

[02:59](javascript:;)These disks have the same maximum IOPS as SSD persistent disks and lower IOPS per gigabyte.

[03:06](javascript:;)For most VM shapes, except very large ones, this disk type offers performance levels suitable

[03:13](javascript:;)for most general-purpose applications at a price point between that of standard and performance persistent disks.

[03:21](javascript:;)Extreme persistent disks are zonal persistent disks also backed by solid-state drives.

[03:26](javascript:;)Extreme persistent disks are designed for high-end database workloads, providing consistently high performance for both random access workloads and bulk throughput.

[03:38](javascript:;)Unlike other disk types, you can provision your desired IOPS.

[03:42](javascript:;)By default, Compute Engine encrypts all data at rest.

[03:47](javascript:;)Google Cloud handles and manages this encryption for you without any additional actions on your part.

[03:55](javascript:;)However, if you wanted to control and manage this encryption yourself, you can either use Cloud Key Management Service to create and manage

[04:02](javascript:;)key encryption keys (which is known as customer-managed encryption keys) or create and manage your own key encryption keys (known as customer-supplied encryption keys).

[04:14](javascript:;)Now, local SSDs are different from persistent disks in that they are physically attached to the virtual machine.

[04:22](javascript:;)Therefore, these disks are ephemeral but provide very high IOPS.

[04:26](javascript:;)For up-to-date numbers I recommend referring to the documentation in the course resources for this module.

[04:32](javascript:;)Currently, local SSDs are 375 gigabytes in size and you can attach up to to 24 local SSD partitions for a total of 9 terabytes per instance.

[04:44](javascript:;)Data on these disks will survive a reset but not a VM stop or terminate, because these disks can’t be reattached to a different VM.

[04:53](javascript:;)You also have the option of using a RAM disk.

[04:56](javascript:;)You can simply use tmpfs if you want to store data in memory.

[05:00](javascript:;)This will be the fastest type of performance available if you need small data structures.

[05:05](javascript:;)I recommend a high-memory virtual machine if you need to take advantage of such features, along with a persistent disk to back up the RAM disk data.

[05:13](javascript:;)In summary, you have several different disk options.

[05:16](javascript:;)Persistent disks can be rebooted and snapshotted, but local SSDs and RAM disks are ephemeral.

[05:24](javascript:;)I recommend choosing a persistent HDD disk when you don't need performance but just need capacity.

[05:31](javascript:;)If you have high performance needs, start looking at the SSD options.

[05:35](javascript:;)The persistent disks offer data redundancy because the data on each persistent disk is distributed across several physical disks.

[05:44](javascript:;)Local SSDs provide even higher performance, but without the data redundancy.

[05:49](javascript:;)Finally, RAM disks are very volatile but they provide the highest performance.

[05:55](javascript:;)Now, just as there is a limit on how many Local SSDs you can attach to a

[05:58](javascript:;)VM, there is also a limit on how many persistent disks you can attach to a VM.

[06:05](javascript:;)As illustrated in this table, this limit depends on the machine type.

[06:10](javascript:;)For the Shared-core machine type, you can attach up to 16 disks.

[06:15](javascript:;)For the Standard, High Memory, High-CPU, Memory-optimized, and Compute-optimized machine types, you can attach up to 128 disks.

[06:23](javascript:;)So you can create massive amounts of capacity for a single host.

[06:28](javascript:;)Now remember that little nuance when I told you about how throughput is limited by the number of cores that you have?

[06:34](javascript:;)That throughput also shares the same bandwidth with Disk IO.

[06:37](javascript:;)So if you plan on having a large amount of Disk IO throughput, it will also compete with any network egress or ingress throughput.

[06:46](javascript:;)So remember that, especially if you will be increasing the number of drives attached to a virtual machine.

[06:53](javascript:;)There are many differences between a physical hard disk in a computer and a persistent disk, which is essentially a virtual networked device.

[07:01](javascript:;)First of all, if you remember with normal computer hardware disks, you have to partition them.

[07:07](javascript:;)Essentially, you have a drive and you’re carving up a section for the operating system to get its own capacity.

[07:13](javascript:;)If you want to grow it, you have to repartition, and if you want to make changes you might even have to reformat.

[07:19](javascript:;)If you want redundancy, you might create a redundant disk array, and if you want encryption, you need to encrypt files before writing them to the disk.

[07:28](javascript:;)With cloud persistent disks, things are very different because all that management is handled for you on the backend.

[07:35](javascript:;)You can simply grow disks and resize the file system because disks are virtual networked devices.

[07:42](javascript:;)Redundancy and snapshot services are built in and disks are automatically encrypted.

[07:47](javascript:;)You can even use your own keys, and that will ensure that no party can get to the data except you.

Common Compute Engine Actions-

[00:00](javascript:;)Now that we have covered all the different compute image and disk options, let's look at some common actions that you can perform with Compute Engine.

[00:07](javascript:;)Every VM instance stores its metadata on a metadata server.

[00:12](javascript:;)The metadata server is particularly useful in combination with startup and shutdown scripts because you

[00:18](javascript:;)can use the metadata server to programmatically get unique information about an instance without additional authorization.

[00:25](javascript:;)For example, you can write a startup script that gets the metadata key value pair for

[00:30](javascript:;)an instance's external IP address and use that IP address in your script to setup a database.

[00:36](javascript:;)Because the default metadata keys are the same on every instance, you can reuse your script without

[00:42](javascript:;)having to update it for each instance, this helps you create less brittle code for your applications.

[00:49](javascript:;)Storing and retrieving instance metadata is a very common Compute Engine action.

[00:54](javascript:;)I recommend storing these startup and shutdown scripts in Cloud Storage as you will explore in the upcoming lab of this module.

[01:01](javascript:;)Another common action is to move an instance to a new zone.

[01:04](javascript:;)For example, you might do so for geographical reasons or because a zone is being deprecated.

[01:11](javascript:;)You can move a VM even if one of the following scenarios apply: The VM instance

[01:15](javascript:;)is in a TERMINATED state, or The VM instance is a Shielded VM that uses UEFI firmware.

[01:22](javascript:;)If you move your instance within the same region, you can automate the move by using the gcloud compute instances move command.

[01:30](javascript:;)To move your VM, you must shut down the VM, move it to the destination zone or region, and then restart it.

[01:36](javascript:;)After you move your VM, update any references that you have to the original resource, such as any target VMs

[01:42](javascript:;)or target pools that point to the earlier VM During the move, some server-generated properties of your VM and disks change.

[01:52](javascript:;)If we move your instance to a different region, you need to manually do so by following the process outlined here.

[01:57](javascript:;)This involves making a snapshot of all persistent disks and creating new disks in the destination zone from that snapshot.

[02:04](javascript:;)Next, you create a new VM in the destination zone and attach the new persistent disks, assign a static IP, and update any references to the VM.

[02:14](javascript:;)Finally, you delete the original VM, its disks and the snapshot.

[02:18](javascript:;)Speaking of snapshots, let's take a closer look at these.

[02:22](javascript:;)Snapshots have many use cases.

[02:24](javascript:;)For example, they can be used to backup critical data into a durable storage solution to meet application, availability, and recovery requirements.

[02:33](javascript:;)These snapshots are stored in Cloud Storage, which is covered later.

[02:37](javascript:;)Snapshots can also be used to micro data between zones.

[02:41](javascript:;)I just discussed this when going over the manual process of moving an instance between two

[02:45](javascript:;)regions, but this can also be used to simply transfer data from one zone to another.

[02:51](javascript:;)For example, you might want to minimize latency by migrating data to a drive that can be locally attached in the zone where it is used.

[02:59](javascript:;)Which brings me to another snapshot use case of transferring data to a different disk type.

[03:04](javascript:;)For example, if you want to improve disk performance, you could use a snapshot to transfer data from a standard ECD persistent disk to a SSD persistent disk.

[03:14](javascript:;)Now that I've covered some of these snapshot use cases, let's explore the concept of a disk snapshot.

[03:21](javascript:;)First of all, this slide is titled persistent disk snapshots because snapshots are available only to persistent disks and not to local SSDs.

[03:31](javascript:;)Snapshots are different from public images and custom images which are used primarily to create instances or

[03:37](javascript:;)configure instance templates, in that snapshots are useful for periodic backup of the data on your persistent disks.

[03:46](javascript:;)Snapshots are incremental and automatically compressed, so you can create regular snapshots on a persistent disk faster

[03:52](javascript:;)and at a much lower cost than if you regularly created a full image of the disk.

[03:59](javascript:;)As we saw with the previous examples, snapshots can be restored to a new persistent disk, allowing for a move to a new zone.

[04:07](javascript:;)To create a persistent disk snapshot, see the link section of this video.

[04:13](javascript:;)Another common Compute Engine action is to resize your persistent disk.

[04:17](javascript:;)The added benefit of increasing storage capacity is to improve I/O performance.

[04:23](javascript:;)This can be achieved while the disk is attached to a running VM without having to create a snapshot.

[04:28](javascript:;)Now, while you can grow disk and size, you can never shrink them.

[04:32](javascript:;)So keep this in mind.

Lab Intro: Working with Virtual Machines-

[00:00](javascript:;)Let's get started with the second lab of this module.

[00:03](javascript:;)In this lab, you will be setting up an application server.

[00:06](javascript:;)Now, this example happens to be a gaming application, but it applies to many other use cases.

[00:12](javascript:;)You will configure the VM and also add capacity for a production gaming system, and you will build the infrastructure that you need for production activities.

[00:21](javascript:;)These include backups and graceful shutdown and restart services.

Working with Virtual Machines- Lab Needed

Lab Review: Working with Virtual Machines-

[00:00](javascript:;)In this lab, you created a customized Virtual Machine instance by installing base software which was a headless Java runtime environment and application software specifically a Minecraft Game Server.

[00:12](javascript:;)You customize the VM by preparing and attaching a high-speed SSD and you've reserved a static external IP address so that the address will remain consistent.

[00:22](javascript:;)Using that IP address, you then verify the availability of the gaming server online.

[00:27](javascript:;)Next, you set up a backup system to backup the service data to a Cloud Storage bucket, and then you tested that backup system.

[00:35](javascript:;)You then automated backups using cron.

[00:38](javascript:;)Finally, you set up maintenance scripts using metadata for graceful startup and shutdown off the server.

[00:44](javascript:;)Many of these techniques including these script automation can be adapted to administration of production servers in any application.

[00:52](javascript:;)You can stay for a lab walk-through.

[00:54](javascript:;)But remember that GCPs user interface can change.

[00:57](javascript:;)So your environment might look slightly different.

[01:00](javascript:;)So here I am in the VM Instances page.

[01:03](javascript:;)Let's go ahead and create our instance.

[01:05](javascript:;)We're going to use the same properties that are provided to us, properties and values in the lab.

[01:11](javascript:;)So I'm going to call this the mc server for our Minecraft Server.

[01:14](javascript:;)We're going to place this in the US Central one, a zone.

[01:17](javascript:;)We're going to modify the access scopes for this.

[01:22](javascript:;)So I'm going to set axis for each API.

[01:25](javascript:;)I'm going to modify for storage that besides just read only, I want read write.

[01:32](javascript:;)This is going to allow the VM instance to write to the Cloud Storage bucket that we're going to create later on.

[01:37](javascript:;)Now we're also going to modify the disk of this instance.

[01:42](javascript:;)So let's expand the option down here, and under disks we're going to add a new disk.

[01:53](javascript:;)We're going to call this the Minecraft disk, and we're going to make that an SSD persistent disk.

[02:06](javascript:;)It's going to be blank so no source.

[02:10](javascript:;)Fifty gigabytes is more than enough for what we're trying to do.

[02:14](javascript:;)I'm going to leave the encryption as Google managed key.

[02:18](javascript:;)So let me click done and this is going to create that disk and automatic attach it to the VM.

[02:24](javascript:;)Now under networking, we're also going to add a network tag.

[02:31](javascript:;)This is going to then allow us to locate specific firewall rules, we call that Minecraft Server.

[02:37](javascript:;)On the network interface I'm going to click on the pencil icon here to edit.

[02:43](javascript:;)We are leaving the internal IP as is but for the external IP, we're actually

[02:47](javascript:;)going to create an IP address which means that we are reserving a static IP address.

[02:53](javascript:;)This is going to make sure that these IP address is not ephemeral and doesn't change.

[02:57](javascript:;)So I'll just give it a name and I click reserve and then we're going

[03:03](javascript:;)to click done once that is reserved and from there we're going to create this instance.

[03:09](javascript:;)It's done and then create.

[03:12](javascript:;)Now once the instance is up and running we're going to have to prepare the data disks.

[03:17](javascript:;)So we're not going to create a directory, format and mount the disk.

[03:22](javascript:;)I don't need this tab over here so I can close that.

[03:26](javascript:;)We're going to wait for the instance there it is.

[03:28](javascript:;)So the SSH to the instance.

[03:33](javascript:;)I'm going to start by creating a director that serves as the amount point for the data off the disk.

[03:40](javascript:;)For that I'm just going to use the command that's provided in the lab and then we're going to format the disk itself.

[03:48](javascript:;)So we're just going to wait for that SSH connection to be established.

[03:54](javascript:;)This is allowed because the default network has a default firewall rule for SSH.

[04:00](javascript:;)So let me go ahead and run that, and then we're going to format the disk.

[04:12](javascript:;)Great.

[04:12](javascript:;)Now we're going to mount it, and this is not going to display any outputs, so don't be surprised about that.

[04:21](javascript:;)There's a checkpoint in the lab.

[04:24](javascript:;)So you can check your progress, worked for me.

[04:27](javascript:;)So I'm going to move on to task three and now install and run the application and the micro server itself runs on top of the Java virtual machine.

[04:35](javascript:;)So we do require the Java Runtime Environment, or JRE, to run.

[04:41](javascript:;)But we don't need the user interface.

[04:43](javascript:;)So we're just going to install actually headless version and that's going to reduce a lot of the resource usage

[04:47](javascript:;)on that machine which will ensure that the Minecraft Server has enough room to expand its own resource usage if needed.

[04:55](javascript:;)So let me go ahead and start by updating the repository.

[05:04](javascript:;)Then I'm going to install that headless JRE, and after that I'm going to navigate to the directory where we mounted that persistent disk.

[05:22](javascript:;)Into that we're then going to download the Minecraft jar file.

[05:30](javascript:;)So we navigate into that under command.

[05:35](javascript:;)You can see it's downloading and the lab manual also provides information on the download page itself.

[05:42](javascript:;)So you can read more about where this comes from.

[05:44](javascript:;)There are also lots of instructions actually in there on how to set this up on a Linux machine.

[05:48](javascript:;)So if you wanted to customize this, I definitely recommend referring to that link.

[05:52](javascript:;)So let's go ahead and initialize the Minecraft Server.

[06:01](javascript:;)Run that command and it's telling us that this is not going to run unless we agree to the end-user licensing agreement.

[06:11](javascript:;)So we need to do that now.

[06:13](javascript:;)Let me just check my progress.

[06:15](javascript:;)Make sure that the JRE installation and Minecraft server installation worked out and I got a green check in my lab.

[06:21](javascript:;)So let's look at the files that were created to identify where this license agreement is, and there it is.

[06:31](javascript:;)We can see it right there.

[06:33](javascript:;)So let me use nano to edit that now.

[06:39](javascript:;)All we really have to do is we have to change this last line, instead of saying false, we just have to agree to it by setting this to true.

[06:47](javascript:;)So let me change that and then we're going to click Control O to write that to that filename hit Enter and then Control X to come back out.

[07:02](javascript:;)So we're not going to try to restart the Minecraft Server yet.

[07:06](javascript:;)We're going to use a different technique in a second.

[07:08](javascript:;)What we're going to do next is we're going to create a virtual terminal screen to start that server, and to do that we're going to install screen.

[07:16](javascript:;)So let's grab that command from the lab instructions.

[07:22](javascript:;)It seems like it was already actually installed.

[07:26](javascript:;)Then we're going to go ahead and start that now using the screen command.

[07:30](javascript:;)So let's run that and this might take a while now but it's going to establish the whole environment for us.

[07:45](javascript:;)So we can see here it's preparing the level world.

[07:47](javascript:;)It's loading some recipe.

[07:48](javascript:;)So these are all now very specific commands in regards to the gaming application that we're installing

[07:54](javascript:;)here and we're going to wait for this to complete before detaching from this and moving on.

[08:03](javascript:;)So we can see that the spawn area here has been completed.

[08:07](javascript:;)We could not detached from this, but one thing I want to point out that we're going to have to

[08:11](javascript:;)do next is when this whole thing started it told us which port it is going to do that for.

[08:18](javascript:;)So the port is right here and we're going to have to create a firewall rule in a second to actually allow client traffic to that port.

[08:27](javascript:;)So we can now detach from this.

[08:30](javascript:;)So we're going to just use Control A and Control D. To get out of here.

[08:35](javascript:;)There is a command if you wanted to reattach to the terminal, we're not going to do that.

[08:39](javascript:;)So I'm just going to exit out of here and we're now going to allow Cloud traffic.

[08:45](javascript:;)So for that, we need to create a firewall rule and we're going to use the

[08:49](javascript:;)network tag that we created which we can display by going to Columns and then Network tags.

[08:55](javascript:;)We can see that Minecraft Server was the network tag.

[08:57](javascript:;)So let's do that.

[08:59](javascript:;)I'm going to navigate to VPC network and specifically Firewall rules.

[09:06](javascript:;)I'm going to give a new firewall rule the name of Minecraft rule.

[09:14](javascript:;)It's going to be on the default network.

[09:17](javascript:;)Could this be the only network we have right now?

[09:19](javascript:;)For specified target tags, we're now going to define Minecraft Server so only apply to the instances that have that tag.

[09:28](javascript:;)So let me define the IP ranges as from anywhere.

[09:37](javascript:;)Now specifically for the protocol that's TCP, and then that port was 25565.

[09:48](javascript:;)Then I'm going to go ahead and click Create.

[09:53](javascript:;)Once it's up and running we're going to verify availability of the server.

[10:02](javascript:;)So I can already start navigating back and I'll monitor the process up here in the notification pane.

[10:08](javascript:;)I'm going back to Compute Engine and we have the external IP address here.

[10:15](javascript:;)We're now going to use a couple of different ways to verify that this is running.

[10:20](javascript:;)Note that we can't click on it because we didn't enable HTTP, that would have been TCP for port 80.

[10:27](javascript:;)In the lab instructions, we have listed a website and we also currently have a Chrome extension there, have that Chrome extension actually right up here.

[10:35](javascript:;)So let's try that.

[10:36](javascript:;)I'm going to go to Options, change the IP address that is in here.

[10:44](javascript:;)Save that, and then we're going to try to verify.

[10:52](javascript:;)I can change this through my Minecraft Server, save those changes and then we're going to keep an eye on here to see if this is coming up.

[11:06](javascript:;)Alternative also we could use any of the websites that are listed in here.

[11:12](javascript:;)Since these are third-party tools sometimes they don't work.

[11:16](javascript:;)So that's definitely something to keep in mind.

[11:18](javascript:;)I think that's actually what's going on right now.

[11:21](javascript:;)With this extension it doesn't seem to want to display this to us right now.

[11:27](javascript:;)But if I check the box in the lab instructions itself, it's telling me that everything is tracked correctly.

[11:33](javascript:;)So we've done all the work.

[11:34](javascript:;)It's just that sometimes, again, these third-party tools that we're using to test the status may not always work.

[11:42](javascript:;)There is another one that I can try really fast.

[11:50](javascript:;)We could grab the external IP address and copy it in there.

[11:58](javascript:;)Get the service status that way.

[12:01](javascript:;)It is telling us that it does have it, so it is up and running

[12:05](javascript:;)currently has no players in it, and it tells us the exact version that we're running.

[12:10](javascript:;)So clearly it is working for this page, just not for the Chrome extension right now.

[12:16](javascript:;)All right.

[12:16](javascript:;)So then let's move on.

[12:17](javascript:;)What we're going to do now is, these services up and running but now we want to actually

[12:21](javascript:;)scheduled some regular backups, have some maintenance around the server so that we plan for the long term.

[12:29](javascript:;)So what I can do now is I can SSH back into the server.

[12:35](javascript:;)Since I allowed for read write access to Cloud Storage, I can actually directly create a bucket now through my server here similarly as you would from Cloud Shell.

[12:47](javascript:;)So the first thing I'm going to do is I'm just going to define my own bucket name, and store it in an environment variable.

[12:55](javascript:;)So here we go, export your bucket name.

[12:58](javascript:;)You want to use something that's globally unique.

[13:01](javascript:;)So one thing we could do is we could take our project ID.

[13:05](javascript:;)You take that right here, and you go back to that server and paste it in there.

[13:18](javascript:;)Whenever you create a an environment variable, you want to run the echo command to make sure that you created it correctly.

[13:22](javascript:;)Here we can see that worked.

[13:25](javascript:;)Now I can use the gsutil command specifically MB for make bucket for Google Cloud Storage, and then use

[13:33](javascript:;)that unique part that I just entered and just append Minecraft backup so that I also know what this is.

[13:42](javascript:;)So this becomes a little bit more readable.

[13:44](javascript:;)Great.

[13:44](javascript:;)So there it is. I could also know verify by the way that it is created in my project.

[13:49](javascript:;)I could go to the navigation menu, and if we go to storage we'll be able to see our bucket right here.

[13:58](javascript:;)We could have also just created this way.

[14:00](javascript:;)But this way we now have everything stored that is the variable in here, and then going forward we can do all of the backup right through the VM.

[14:10](javascript:;)So let's go ahead, and create a backup script.

[14:13](javascript:;)I'm just going to navigate to the home directory that we have within Minecraft, and we're going to just create a new script using nano.

[14:24](javascript:;)I'm going to paste the script that we already have in here which has the screen command, and then talks about the backups.

[14:31](javascript:;)So let me paste this in there, and then we're going to press Control O, and then enter to save and control X to come back.

[14:41](javascript:;)So this script saves the current state of the server's world and pauses a service odyssey functionality.

[14:46](javascript:;)Then it's going to backup the service world data directory and place its content in a Timestamp directory in the Cloud Storage bucket.

[14:53](javascript:;)After the script, I've finished his backup the data it resumes odyssey saving on the Minecraft Server.

[14:58](javascript:;)Now we got to make sure that this is actually executable.

[15:02](javascript:;)So let's run the following command, and now we can go and test this.

[15:09](javascript:;)So let's actually run the backup script so there we can see that we are copying some files.

[15:20](javascript:;)Let's verify that.

[15:20](javascript:;)So I'm going to now navigate into my Cloud Storage bucket that I already have here.

[15:26](javascript:;)If I open that, we can now see a folder in there, and I could dig further into there to get more information about the world.

[15:33](javascript:;)So clearly we can see that the backup is working for us.

[15:38](javascript:;)We can also now schedule the backup to run in and more automated fashion.

[15:44](javascript:;)So I'm going to go back to my SSH session, run the pseudo crontab command.

[15:52](javascript:;)Now we want to choose nano in this case, it does tell us it's easiest but you do have other options available if those are more comfortable.

[16:02](javascript:;)At the bottom, we're now going to define how often this runs.

[16:08](javascript:;)This is going to tell it to run the backup every four hours.

[16:13](javascript:;)There's documentation that you can look into and how to define this, but in this case, that's more than enough for what we're trying to achieve.

[16:19](javascript:;)So let's save that file and get back out.

[16:24](javascript:;)This is going to create a lot of backups.

[16:26](javascript:;)I mean about 300 a month.

[16:28](javascript:;)So maybe you want to look into regular deleting those Cloud Storage does offer Object Lifecycle Management features that

[16:35](javascript:;)let you set the time to live for objects and even archive older objects to a different Storage class.

[16:42](javascript:;)You'll learn more about that in the next course of this series when we talk about Cloud Storage.

[16:50](javascript:;)I'm just going to go ahead and check my progress.

[16:52](javascript:;)In my lab looks like everything worked.

[16:55](javascript:;)The last thing we're going to do is now perform some maintenance.

[16:59](javascript:;)So specifically when we shut down and restart that certain actions happen.

[17:04](javascript:;)So let me run the pseudo screen command, and then I'm going to go and actually stop this instance.

[17:18](javascript:;)So I'm going to go navigate to Compute Engine, click on the server so select it and click stop.

[17:32](javascript:;)It's going to ask us if we sure we want to do that, and yes we're going to stop.

[17:37](javascript:;)Then later if we want to start it back up we can do that.

[17:40](javascript:;)This is also going to log us out of our SSH session obviously.

[17:44](javascript:;)So let's wait for this to stop, and then we're going to automate the server maintenance with some startup and shutdown scripts.

[17:52](javascript:;)So the instance has stopped, am going to click on it now to edit some of the custom metadata.

[17:58](javascript:;)So let me click Edit, and we're going to scroll down to the metadata.

[18:04](javascript:;)Here we go.

[18:08](javascript:;)What we're going to define now is a startup script as well as the shutdown script.

[18:12](javascript:;)We're going to point those to files that we have in Cloud Storage that are publicly available.

[18:18](javascript:;)So the key is going to be startup script URL, and then the value is going to be the location of the file.

[18:28](javascript:;)I can make them bigger to make sure that formatting that correctly.

[18:33](javascript:;)I'll add another item, and we'll do the same for the shutdown script.

[18:36](javascript:;)You can actually navigate yourself to these files if you want to, and you could read more about what exactly happens in these startup and shutdown script.

[18:44](javascript:;)So now I can click Save, and I could restart the service.

[18:51](javascript:;)I did in the meantime while the service was shutting down.

[18:54](javascript:;)I went back to the status page, and you can see that the status as currently says could not resolve so clearly the server is shut down.

[19:05](javascript:;)Now when we restart this, once all the startup script is done running, we can go back and we can verify that this service is indeed now accessible again.

[19:18](javascript:;)Just keep in mind that that might take a while for the actual instance to startup which it is now, and then for the startup script to actually finish.

Quiz: Virtual Machines-

Which statement is true of persistent disks?

Persistent disks are encrypted by default.

Persistent Disks are not physical disks, they are a virtual-networked service. Each persistent disk remains encrypted either with system-defined keys or with customer-supplied keys.

Which statement is true of Virtual Machine Instances in Compute Engine?

In Compute Engine, a VM is a networked service that simulates the features of a computer.

VMs in Compute Engine are a collection of networked services which includes persistent disks that are network-attached. In some cases the Google Cloud VM behaves unlike hardware or other kinds of virtual machines, for example, when a multi-tenant virtual CPU "bursts", using excess capacity beyond the VM spec.

What are sustained use discounts?

Automatic discounts that you get for running specific Compute Engine resources for a significant portion of the billing month

That's correct! Sustained use discounts are automatic discounts that you get for running specific Compute Engine resources (vCPUs, memory, GPU devices) for a significant portion of the billing month. To take advantage of the full 30% discount, create your VM instances on the first day of the month, because discounts reset at the beginning of each month.

Module Review-

[00:00](javascript:;)In this module, we cover the different Compute image and disk options within Compute Engine, along with some common actions.

[00:07](javascript:;)The two labs provided you with real-world applications of most of the topics covered in this course.

[00:13](javascript:;)Remember there are many Compute options to choose from.

[00:16](javascript:;)If a predefined machine type does not meet your needs, you can also customize your own VM and you can even create a sole-tenant node.

[00:23](javascript:;)You can also install different public and custom images on the boot disk of you instances, and you can attach more disks if needed.

Course Review-

[00:00](javascript:;)Thank you for taking the Essential Cloud Infrastructure Foundation course.

[00:04](javascript:;)I hope you have a better understanding of how to architect with Compute Engine, and I also hope

[00:08](javascript:;)that the demos and labs made you feel more comfortable with using the different GCP services that we covered.

[00:15](javascript:;)Next, I recommend enrolling in the Essential Cloud Infrastructure Core Services course of the architecting with Google Compute Engine series.

[00:24](javascript:;)In that course, we start by talking about Cloud IAM, and you will administer identity and access management for resources.

[00:31](javascript:;)Next, we'll cover the different data storage services in GCP, and you will implement some of those services.

[00:38](javascript:;)Then will go into resource management where you will manage and examine billing data of GCP resources.

[00:44](javascript:;)Lastly, we'll talk about resource monitoring, and you will monitor GCP resources using Stackdriver services.

[00:51](javascript:;)Enjoy that course.

Cloud shell provides the following:

* Temporary Compute Engine VM
* Command-line access to the instance via a browser
* 5 GB of persistent disk storage ($HOME dir)
* Pre-installed Cloud SDK and other tools
* gcloud: for working with Compute Engine and many Google Cloud services
* gcloud storage: for working with Cloud Storage
* kubectl: for working with Google Kubernetes Engine and Kubernetes
* bq: for working with BigQuery
* Language support for Java, Go, Python, Node.js, PHP, and Ruby
* Web preview functionality
* Built-in authorization for access to resources and instances

Cloud Shell provides you with which of the following? (Select all that apply).

5 GB of persistent storage (/home)

Built-in authorization for access to resources and instances

Command-line access to a free temporary Compute Engine VM

Google Cloud Marketplace lets you quickly deploy functional software packages by providing pre-defined templates with which Google Cloud service?

Deployment Manager

Deployment Manager is a Google Cloud service that uses templates written in a combination of YAML, python, and Jinja2 to automate the allocation of Google Cloud resources and perform setup tasks. Behind the scenes a virtual machine has been created. A startup script was used to install and configure software, and network Firewall Rules were created to allow traffic to the service.

Google Cloud Virtual Private Cloud (VPC) provides networking functionality to Compute Engine virtual machine (VM) instances, Kubernetes Engine containers, and the App Engine flexible environment. In other words, without a VPC network, you cannot create VM instances, containers, or App Engine applications. Therefore, each Google Cloud project has a **default** network to get you started.

You can think of a VPC network as similar to a physical network, except that it is virtualized within Google Cloud. A VPC network is a global resource that consists of a list of regional virtual subnetworks (subnets) in data centers, all connected by a global wide area network (WAN). VPC networks are logically isolated from each other in Google Cloud.

Routes tell VM instances and the VPC network how to send traffic from an instance to a destination, either inside the network or outside Google Cloud. Each VPC network comes with some default routes to route traffic among its subnets and send traffic from eligible instances to the internet.

Each VPC network implements a distributed virtual firewall that you can configure. Firewall rules allow you to control which packets are allowed to travel to which destinations. Every VPC network has two implied firewall rules that block all incoming connections and allow all outgoing connections.

These firewall rules allow **ICMP**, **RDP**, and **SSH** ingress traffic from anywhere (0.0.0.0/0) and all **TCP**, **UDP**, and **ICMP** traffic within the network (10.128.0.0/9). The **Targets**, **Filters**, **Protocols/ports**, and **Action** columns explain these rules.

**Note:**Without a VPC network, there are no routes!

Without a VPC network, you cannot create VM instances, containers, or App Engine applications.

True

**Note:**As expected, you cannot create a VM instance without a VPC network!

You have been tasked to create an auto mode network with two VM instances. Auto mode networks are easy to set up and use because they automatically create subnets in each region. However, you don't have complete control over the subnets created in your VPC network, including regions and IP address ranges used.

Subnets let you create your own private cloud topology within Google Cloud.

**Note:**The **External IP addresses** for both VM instances are ephemeral. If an instance is stopped, any ephemeral external IP addresses assigned to the instance are released back into the general Compute Engine pool and become available for use by other projects.

When a stopped instance is started again, a new ephemeral external IP address is assigned to the instance. Alternatively, you can reserve a static external IP address, which assigns the address to your project indefinitely until you explicitly release it.

**Note:**You can SSH because of the **allow-ssh** firewall rule, which allows incoming traffic from anywhere (0.0.0.0/0) for **tcp:22**. The SSH connection works seamlessly because Compute Engine generates an SSH key for you and stores it in one of the following locations:

* By default, Compute Engine adds the generated key to project or instance metadata.
* If your account is configured to use OS Login, Compute Engine stores the generated key with your user account.

Alternatively, you can control access to Linux instances by creating SSH keys and editing public SSH key metadata.

You can ping **mynet-eu-vm**'s internal IP because of the **allow-custom** firewall rule.

Which firewall rule allows the ping to mynet-eu-vm's external IP address?

mynetwork-allow-icmp

**Note:**You can ping the external IP address of all VM instances, even though they are in either a different zone or VPC network. This confirms that public access to those instances is only controlled by the **ICMP** firewall rules that you established earlier.

**Note:**You can ping the internal IP address of **mynet-eu-vm** because it is on the same VPC network as the source of the ping (**mynet-us-vm**), even though both VM instances are in separate zones, regions, and continents!

**Note:**This should not work either, as indicated by a 100% packet loss! You cannot ping the internal IP address of **managementnet-us-vm** and **privatenet-us-vm** because they are in separate VPC networks from the source of the ping (**mynet-us-vm**), even though they are all in the same zone.

In VPC Networking lab, you explored the default network and determined that you cannot create VM instances without a VPC network. Thus, you created a new auto mode VPC network with subnets, routes, firewall rules, and two VM instances and tested the connectivity for the VM instances. Because auto mode networks aren't recommended for production, you converted the auto mode network to a custom mode network.

Next, you created two more custom mode VPC networks with firewall rules and VM instances using the Cloud Console and the gcloud command line. Then you tested the connectivity across VPC networks, which worked when pinging external IP addresses but not when pinging internal IP addresses.

VPC networks are by default isolated private networking domains. Therefore, no internal IP address communication is allowed between networks, unless you set up mechanisms such as VPC peering or VPN.

In this lab, you implement Private Google Access and Cloud NAT for a VM instance that doesn't have an external IP address. Then, you verify access to public IP addresses of Google APIs and services and other connections to the internet.

VM instances without external IP addresses are isolated from external networks. Using Cloud NAT, these instances can access the internet for updates and patches, and in some cases, for bootstrapping. As a managed service, Cloud NAT provides high availability without user management and intervention.

**Note:** When instances do not have external IP addresses, they can only be reached by other instances on the network via a managed VPN gateway or via a Cloud IAP tunnel. Cloud IAP enables context-aware access to VMs via SSH and RDP without bastion hosts. To learn more about this, see the blog post [Cloud IAP enables context-aware access to VMs via SSH and RDP without bastion hosts](https://cloud.google.com/blog/products/identity-security/cloud-iap-enables-context-aware-access-to-vms-via-ssh-and-rdp-without-bastion-hosts).

VM instances that have no external IP addresses can use Private Google Access to reach external IP addresses of Google APIs and services. By default, Private Google Access is disabled on a VPC network.

Private Google Access is enabled at the subnet level. When it is enabled, instances in the subnet that only have private IP addresses can send traffic to Google APIs and services through the default route (0.0.0.0/0) with a next hop to the default internet gateway.

Cloud NAT is a regional resource. You can configure it to allow traffic from all ranges of all subnets in a region, from specific subnets in the region only, or from specific primary and secondary CIDR ranges only.

Cloud NAT lets your Compute Engine instances and Kubernetes Engine container pods communicate with the internet using a shared, public IP address. Cloud NAT uses a Cloud NAT gateway to connect your subnets to a Cloud Router, a virtual router that connects to the internet.

**Note:** The Cloud NAT gateway implements outbound NAT, but not inbound NAT. In other words, hosts outside of your VPC network can only respond to connections initiated by your instances; they cannot initiate their own, new connections to your instances via NAT.

As a reminder, Cloud NAT logs are generated for the following sequences:

* When a network connection using NAT is created.
* When a packet is dropped because no port was available for NAT.

You created **vm-internal**, an instance with no external IP address, and connected to it securely using an IAP tunnel. Then you enabled Private Google Access, configured a NAT gateway, and verified that **vm-internal** can access Google APIs and services and other public IP addresses.

VM instances without external IP addresses are isolated from external networks. Using Cloud NAT, these instances can access the internet for updates and patches, and in some cases, for bootstrapping. As a managed service, Cloud NAT provides high availability without user management and intervention.

IAP uses your existing project roles and permissions when you connect to VM instances. By default, instance owners are the only users that have the **IAP Secured Tunnel User** role.

**Note:**Notice that you cannot change the machine type, the CPU platform, or the zone.

You can add network tags and allow specific network traffic from the internet through firewalls. Some properties of a VM are integral to the VM, are established when the VM is created, and cannot be changed. Other properties can be edited.

You can add additional disks and you can also determine whether the boot disk is deleted when the instance is deleted.

Normally the boot disk defaults to being deleted automatically when the instance is deleted. But sometimes you will want to override this behavior. This feature is very important because you cannot create an image from a boot disk when it is attached to a running instance.

So you would need to disable **Delete boot disk when instance is deleted** to enable creating a system image from the boot disk.

**Note:**You cannot convert a non-preemptible instance into a preemptible one. This choice must be made at VM creation. A preemptible instance can be interrupted at any time and is available at a lower cost.

If a VM is stopped for any reason, (for example an outage or a hardware failure) the automatic restart feature will start it back up. Is this the behavior you want? Are your applications idempotent (written to handle a second startup properly)?

During host maintenance, the VM is set for live migration. However, you can have the VM terminated instead of migrated.

If you make changes, they can sometimes take several minutes to be implemented, especially if they involve networking changes like adding firewalls or changing the external IP.

**Note:**When the VM is running, notice that the connection option in the far right column is RDP, not SSH. RDP is the Remote Desktop Protocol. You would need the RDP client installed on your local machine to connect to the Windows desktop.

**Note:**Installing an RDP client on your local machine is outside the scope of this lab and of the class. For this reason, you will not be connecting to the Windows VM during this lab. However, you will step through the usual procedures up to the point of requiring the RDP client. Instructions for connecting to Windows VMs are in the [Connecting to Windows VMs Guide](https://cloud.google.com/compute/docs/instances/windows/connecting-to-windows-instance).

gcloud storage buckets create gs://$YOUR\_BUCKET\_NAME-minecraft-backup

Cloud Storage offers the Object Lifecycle Management feature to set a Time to Live (TTL) for objects, archive older versions of objects, or "downgrade" storage classes of objects to help manage costs.

In Working with Virtual Machines lab, you created a customized virtual machine instance by installing base software (a headless JRE) and application software (a Minecraft game server). You customized the VM by attaching and preparing a high-speed SSD data disk, and you reserved a static external IP so the address would remain consistent. Then you verified availability of the gaming server online.

You set up a backup system to back up the server's data to a Cloud Storage bucket, and you tested the backup system. Then you automated backups using cron. Finally, you set up maintenance scripts using metadata for graceful startup and shutdown of the server.

ESSENTIAL GOOGLE CLOUD INFRASTRUCTURE: CORE SERVICES:

Course Introduction-

[00:02](javascript:;)Hello.

[00:02](javascript:;)I'm Philipp Maier.

[00:05](javascript:;)I'm Mylene Biddle, we're both Course Developers, at Google Cloud and we want to welcome you to Architecting with Compute Engine, a series of three courses.

[00:13](javascript:;)Before we start using all of the different services that Google Cloud Platform, or GCP offers, let's talk about what GCP is.

[00:21](javascript:;)When you look at Google Cloud, you'll see that it's actually part of a much larger ecosystem.

[00:27](javascript:;)This ecosystem consists of open-source software, providers, partners, developers, third-party software, and other Cloud providers.

[00:35](javascript:;)Google is actually a very strong supporter of open-source software.

[00:39](javascript:;)That's right.

[00:39](javascript:;)Now, Google Cloud consists of Chrome, Google devices, Google Maps, Gmail, Google Analytics, G Suite, Google Search, and the Google Cloud Platform.

[00:49](javascript:;)GCP itself is a computing solution platform that really encompasses three core features: infrastructure, platform, and software.

[00:58](javascript:;)This map represents GCP's global infrastructure.

[01:02](javascript:;)As of this recording, GCP's well-provisioned global network connects over 60 zones to over 130 points of presence through a global network of fiber optic cables.

[01:14](javascript:;)And Google is continuously investing in this network, with new regions, points of presence, and subsea cable investments.

[01:21](javascript:;)On top of this infrastructure, GCP uses state of the art software-defined, networking and distributed systems of technologies to host and deliver your services around the world.

[01:31](javascript:;)These technologies are represented by a suite of Cloud-based products and services that is continuously expanding.

[01:38](javascript:;)Now, it's important to understand that there is usually more than one solution for a task or application in GCP.

[01:45](javascript:;)To better understand this, let's look at a solution continuum.

[01:49](javascript:;)Google Cloud Platform spans from infrastructure as a service, or IaaS, to software as a service, or SaaS.

[01:57](javascript:;)You really can build applications on GCP for the web or mobile that are global,

[02:02](javascript:;)auto-scaling, and assistive, and that provide services where the infrastructure is completely invisible to the user.

[02:08](javascript:;)It is not just that Google has opened the infrastructure that powers applications like Search, Gmail, Google Maps, and G Suite.

[02:15](javascript:;)Google has opened all of the services that make these products possible and packaged them for your use.

[02:21](javascript:;)Alternative solutions are possible.

[02:23](javascript:;)For example, you could start up your own VM in Google Compute Engine, install open-source MySQL on

[02:28](javascript:;)it and run it just like a MySQL database on your own computer in a data center.

[02:35](javascript:;)Or you could use the Cloud SQL service, which provides a MySQL instance and handles operational work

[02:41](javascript:;)like backups and security patching for you using the same services Google does to automate backups and patches.

[02:49](javascript:;)You could even move to a NoSQL database that is auto-scaling and serverless so that growth

[02:55](javascript:;)no longer requires adding server instances or possibly changing the design to handle the new capacity.

[03:02](javascript:;)This series of courses focuses on the infrastructure.

[03:05](javascript:;)An IT infrastructure is like a city infrastructure.

[03:09](javascript:;)The infrastructure is the basic underlying framework of fundamental facilities and systems, such as transport, communications, power, water, fuel, and other essential services.

[03:20](javascript:;)The people in the city are like users, and the cars and bikes, and buildings in the city are like applications.

[03:28](javascript:;)Everything that goes into creating and supporting those applications for the users is the infrastructure.

[03:33](javascript:;)The purpose of this course is to explore as efficiently and clearly as possible the infrastructure services provided by GCP.

[03:42](javascript:;)You should become familiar enough with the infrastructure services that you will know what services do and how to use them.

[03:49](javascript:;)We won't go into very deep dive case studies on specific vertical applications.

[03:54](javascript:;)But you'll know enough to put all the building blocks together to build your own solution.

[03:59](javascript:;)Now, GCP offers a range of compute services.

[04:03](javascript:;)The service that might be most familiar to newcomers is Compute Engine, which lets you run virtual machines on-demand in the Cloud.

[04:11](javascript:;)It's Google Cloud's infrastructure as a service solution.

[04:14](javascript:;)It provides maximum flexibility for people who prefer to managed server instances themselves.

[04:20](javascript:;)Google Kubernetes Engine lets you run containerized applications on a cloud environment that Google manages for you under your administrative control.

[04:29](javascript:;)Think of containerization as a way to package code that's designed to be highly portable and to use resources very efficiently.

[04:37](javascript:;)And think of Kubernetes as a way to orchestrate code in containers.

[04:41](javascript:;)App Engine is GCP's fully managed platform as a service framework.

[04:46](javascript:;)That means it's a way to run code in the cloud without having to worry about infrastructure.

[04:52](javascript:;)You just focus on your code and let Google deal with all the provisioning and resource management.

[04:57](javascript:;)You can learn a lot more about App Engine in the "Developing Applications with Google Cloud Platform" course series.

[05:04](javascript:;)Cloud Functions is a completely serverless execution environment or functions as a service.

[05:09](javascript:;)It executes your code in response to events, whether those events occur once a day or many times per second.

[05:16](javascript:;)Google scales resources as required, but you only pay for the service while your code runs.

[05:21](javascript:;)The "Developing Applications with Google Cloud" course series also discusses Cloud Functions.

[05:27](javascript:;)Cloud Run, a managed compute platform that lets you run stateless containers via web requests or Pub/Sub events.

[05:33](javascript:;)Cloud Run is serverless.

[05:34](javascript:;)That means it removes all infrastructure management tasks so you can focus on developing applications.

[05:41](javascript:;)It is built on Knative, an open API and runtime environment built on Kubernetes that gives you freedom to move your workloads across different environments and platforms.

[05:50](javascript:;)It can be fully managed on Google Cloud, on Google Kubernetes Engine, or anywhere Knative runs.

[05:57](javascript:;)Cloud Run is fast.

[05:58](javascript:;)It can automatically scale up and down from zero almost instantaneously, and it charges you only for the

[06:04](javascript:;)resources you use calculated down to the nearest 100 milliseconds, so you‘ll never pay for your over-provisioned resources.

[06:13](javascript:;)In this series of courses, In this series of courses, Compute Engine will be our main focus.

[06:18](javascript:;)The Architecting with Google Compute Engine courses are part of the Cloud Infrastructure learning path.

[06:23](javascript:;)This path is designed for IT professionals who are responsible for implementing, deploying, migrating, and maintaining applications in the cloud.

[06:32](javascript:;)The prerequisite for these courses is the Google Cloud Platform Fundamentals: Core Infrastructure course, which you can find in the link section for this video.

[06:42](javascript:;)The Architecting with Google Compute Engine series consists of three courses.

[06:47](javascript:;)Essential Cloud Infrastructure: Foundation is the first course of the Architecting with Compute Engine series.

[06:54](javascript:;)In that course, we start by introducing you to GCP and how to interact with the GCP Console and Cloud Shell.

[07:01](javascript:;)Next, we'll get into virtual networks and you will create VPC networks and other networking objects.

[07:07](javascript:;)Then we'll take a deep dive into virtual machines, and you will create virtual machines using Compute Engine.

[07:13](javascript:;)Essential Cloud Infrastructure: Core Services is the second course of this series.

[07:18](javascript:;)In that course, we start by talking about Cloud IAM and you will administer Identity and Access Management for resources.

[07:25](javascript:;)Next, we'll cover the different data storage services in GCP, and you will implement some of those services.

[07:31](javascript:;)Then we'll go over resource management, where you will manage and examine billing of GCP resources.

[07:37](javascript:;)Lastly, we'll talk about resource monitoring and you will monitor GCP resources using Stackdriver services.

[07:44](javascript:;)Elastic Cloud Infrastructure: Scaling, and Automation, is the last course of the series.

[07:49](javascript:;)In that course, we start by going over the different options to interconnect networks to enable you to connect your infrastructure to GCP.

[07:57](javascript:;)Next, we'll go over GCP is load balancing and auto-scaling services.

[08:01](javascript:;)Would you will get to explore directly.

[08:04](javascript:;)Then we'll cover infrastructure automation services like Terraform so that you can automate the development of GCP infrastructure services.

[08:12](javascript:;)Lastly, we'll talk about other managed services that you might want to leverage in GCP.

[08:17](javascript:;)Now, our goal for you is to remember and understand the different GCP services and features, and

[08:23](javascript:;)also be able to apply your knowledge, analyze requirements, evaluate different options, and create your own services.

[08:30](javascript:;)That's why these courses include interactive hands-on maps through the Qwiklabs platform.

[08:35](javascript:;)Qwiklabs provisions you with a Google account and credentials, so you can access the GCP console for each lab at no cost.

IDENTITY AND ACCESS MANAGEMENT (IAM):

Module Overview-

[00:00](javascript:;)In this module we cover Cloud Identity and Access Management or Cloud IAM.

[00:05](javascript:;)Cloud IAM is a sophisticated system built on top of email-like address names, job type roles in granular permissions.

[00:13](javascript:;)If you're familiar with IAM from other implementations look for the differences that Google has implemented to make IAM easier to administer and more secure.

[00:23](javascript:;)I will start by introducing Cloud IAM from a high-level perspective.

[00:27](javascript:;)We will then dive into each of the components within Cloud IAM which are organizations, roles, members and service accounts.

[00:36](javascript:;)I will also introduce some best practices to help you apply these concepts in your day-to-day work.

[00:42](javascript:;)Finally, you will gain firsthand experience with Cloud IAM through a Lab.

[00:47](javascript:;)Let's get started with an overview of Cloud Identity and Access Management.

Identity and Access Management-

[00:00](javascript:;)Person: So what is Identity Access Management?

[00:04](javascript:;)It is a way of identifying who can do what on which resource.

[00:08](javascript:;)The who can be a person, group or application.

[00:12](javascript:;)The what refers to specific privileges or actions, and the resource could be any Google Cloud service.

[00:19](javascript:;)For example, I could give you the privilege or role of compute viewer.

[00:24](javascript:;)This proves you with read-only access to get and list compute engine resources without being able to read the data stored on them.

[00:32](javascript:;)Cloud IAM is composed of different objects as shown on the slide.

[00:37](javascript:;)We are going to cover each of these in this module.

[00:40](javascript:;)To get a better understanding of where these fit in, let's look at Cloud IAM policies and the Cloud IAM resource hierarchy.

[00:47](javascript:;)Google Cloud resources are organized hierarchically as shown in this tree structure.

[00:53](javascript:;)The organization node is the root node in this hierarchy.

[00:56](javascript:;)Folders are the children of the organization.

[01:00](javascript:;)Projects are the children of folders, and individual resources are the children of projects.

[01:06](javascript:;)Each resource has exactly one parent.

[01:09](javascript:;)The organization resource represents your company.

[01:13](javascript:;)Cloud IAM roles granted by this level are inherited by all resources under the organization.

[01:19](javascript:;)The folder resource could represent your department.

[01:22](javascript:;)Cloud IAM roles granted at this level are inherited by all resources that the folder contains.

[01:27](javascript:;)Projects represent a trust boundary within your company.

[01:31](javascript:;)Services within the same project have the same default level of trust.

Organization-

[00:00](javascript:;)Let's learn more about the organization node.

[00:03](javascript:;)As I mentioned earlier, the organization resource is the root node in the GCP resource hierarchy.

[00:09](javascript:;)This node has many roles, like the organization admin.

[00:13](javascript:;)The organization admin provides a user like Bob, with access to administer all resources belonging to his organization, which is useful for auditing.

[00:22](javascript:;)There is also a project creator role, which allows a user like Alice, to create projects within her organization.

[00:28](javascript:;)I am showing the project creator role here because it can also be applied at the organization level, which would then be inherited by all the projects within the organization.

[00:38](javascript:;)The organization resource is closely associated with a G Suite or Cloud Identity Account.

[00:43](javascript:;)When a user with a G Suite or Cloud Identity Account creates a GCP project an organization resource is automatically provisioned for them.

[00:51](javascript:;)Then Google Cloud communicates its availability to the G Suite or Cloud Identity super admins.

[00:57](javascript:;)These super admin accounts, should be used very carefully because they have a lot of control over your organization and all the resources underneath it.

[01:06](javascript:;)The G Suite or Cloud Identity super administrators and the GCP organization admin are key roles during the setup process and for lifecycle control, for the organization resource.

[01:16](javascript:;)The two roles are generally assigned to different users or groups, although this depends on the organization structure and needs.

[01:23](javascript:;)In the context of GCP organization setup, the G Suite or Cloud Identity super administrator responsibilities are: assign the organization admin role to some

[01:32](javascript:;)users, be a point of contact in case of recovery issues, control the lifecycle of the G Suite or Cloud Identity account and organization resource.

[01:42](javascript:;)The responsibilities of the organization admin role are: define IAM policies, determine the structure of the

[01:49](javascript:;)resource hierarchy, delegate responsibility over critical components such as networking, billing, and resource hierarchy, through IAM roles.

[01:59](javascript:;)Following the principle of least privilege, this role does not include the permission to perform other actions, such as creating folders.

[02:07](javascript:;)To get these permissions, an organization admin must assign additional roles to their account.

[02:12](javascript:;)Let's talk more about folders, because they can be viewed as sub organizations within the organization.

[02:19](javascript:;)Folders provide an additional grouping mechanism and isolation boundary between projects.

[02:24](javascript:;)Folders can be used to model different legal entities, departments, and teams within a company.

[02:30](javascript:;)For example, a first-level of folders can be used to represent the main departments in your organization, like departments x and y.

[02:38](javascript:;)Because folders, can contain projects in other folders, each folder could then include other subfolders to represent different teams, like

[02:46](javascript:;)teams A and B. Each team folder could contain additional subfolders, to represent different applications, like products 1 and 2.

[02:56](javascript:;)Folders allow delegation of administration rights, for example, each head of a department, can be granted full ownership of all GCP resources that belong to their department.

[03:06](javascript:;)Similarly, access to resources can be limited by folder, so users in one department can only access and create GCP resources, within that folder.

[03:16](javascript:;)Let's look at some other resource manager roles, while remembering that policies are inherited from top to bottom.

[03:23](javascript:;)The organization node also has a viewer role.

[03:25](javascript:;)They grants view access to all resources within an organization.

[03:29](javascript:;)The folder node has multiple roles that mimic the organizational roles, but are applied to resources within a folder.

[03:36](javascript:;)There is an admin role that provides full control over folders.

[03:39](javascript:;)A creator role, to browse the hierarchy and create folders, and a viewer role, to view folders and projects below a resource.

[03:47](javascript:;)Similarly for projects, there is a creator role that allows a user to create new projects, making that user automatically the owner.

[03:56](javascript:;)There is also a project deleter role that grants deletion privileges for projects.

Roles-

[00:00](javascript:;)Let's talk more about roles which define the can do what on which resource part of Cloud IAM.

[00:06](javascript:;)There are three types of roles in Cloud IAM, basic roles, predefined roles, and custom roles.

[00:14](javascript:;)Basic roles are the original roles that were available in the Cloud console, but they are broad.

[00:19](javascript:;)You apply them to a Google Crowd project, and they affect all resources in that project.

[00:26](javascript:;)In other words, IAM basic roles offer fixed, coarse-grained levels of access.

[00:31](javascript:;)The basic roles are the owner, editor and viewer roles.

[00:35](javascript:;)The owner has full administrative access.

[00:38](javascript:;)This includes the ability to add and remove members and delete projects.

[00:43](javascript:;)The editor role has modify and delete access.

[00:46](javascript:;)This allows the developer to deploy applications and modify or configure its resources.

[00:52](javascript:;)The view role has read only access.

[00:55](javascript:;)Now all of these roles are concentric.

[00:57](javascript:;)That is the owner role includes the permissions of the editor role.

[01:01](javascript:;)And the editor role includes the permissions of the viewer role.

[01:05](javascript:;)There is also a billing administrator role to manage billing and add or remove administrators without the right to change the resources in the project.

[01:14](javascript:;)Each project can have multiple owners, editors, viewers and billing administrators.

[01:21](javascript:;)GCP services, offers their own set of predefined roles, and they define where the roles can be applied.

[01:28](javascript:;)This provides members with granular access to specific GCP resources and prevents unwanted access to other resources.

[01:36](javascript:;)These roles are a collection of permissions, because to do any meaningful operations, you usually need more than one permission.

[01:46](javascript:;)For example, as shown here, a group of users is granted the instance admin role on project a.

[01:52](javascript:;)This provides the users of that group with all the Compute Engine permissions listed on the right and even more.

[01:59](javascript:;)Grouping these permissions into a role makes them easier to manage.

[02:03](javascript:;)The permissions themselves are classes and methods in the API's.

[02:08](javascript:;)For example, compute instance start can be broken down into the service, resource and verb.

[02:14](javascript:;)That mean that the permission is used to start a stopped Compute Engine instance.

[02:19](javascript:;)These permissions usually line with the actions corresponding REST API.

[02:26](javascript:;)Compute Engine has several predefined IAM roles.

[02:29](javascript:;)Let's look at three of those.

[02:31](javascript:;)The Compute Admin role provides full control of all Compute Engine resources.

[02:36](javascript:;)This includes all permissions that start with compute, which means that every action for any type of Compute Engine resource is permitted.

[02:44](javascript:;)The Network Admin role Contains permissions to create, modify and delete network resources, except for firewall rules and SSL certificates.

[02:54](javascript:;)In other words, the network admin role allows read only access to firewall rules SSL certificates, and instances to view their ephemeral IP addresses.

[03:04](javascript:;)The storage admin role contains permissions to create, modify, and delete disks, images, and snapshots.

[03:11](javascript:;)For example, if your company has someone who manages project images, and you don't want them to have the editor role in the project.

[03:18](javascript:;)Grant their account the storage admin role on that project.

[03:22](javascript:;)For the full list of predefined roles for Compute Engine, see the links section in the slides.

[03:29](javascript:;)Now, roles are meant to represent abstract functions and are customized to line with real jobs.

[03:35](javascript:;)But what if one of those roles does not have enough permissions?

[03:39](javascript:;)Or you need something even finer grained?

[03:42](javascript:;)That's what custom roles permit.

[03:44](javascript:;)A lot of companies use the least privileged model in which each person in your organization is giving the minimal amount of privilege needed to do their job.

[03:54](javascript:;)Let's say you want to define an instance operator role to allow some users to start and stop Compute Engine virtual machines, but not reconfigure them.

[04:03](javascript:;)Custom roles allow you to do that.

Demo: Custom Roles-

[00:00](javascript:;)Let me show you how to create a custom role in GCP.

[00:04](javascript:;)My goal is to create an instance operator role that allows some users to start and stop Compute Engine virtual machines but not reconfigure them.

[00:13](javascript:;)So here I am in the GCP console and I'm going to click the "Navigation"

[00:17](javascript:;)menu to go to "I am an Admin" and specifically actually want to go to "Roles".

[00:28](javascript:;)Here I can see all the different roles that are available.

[00:31](javascript:;)Now, I could select one of these roles and create a role from that selection and then either remove or assign more permissions.

[00:40](javascript:;)You can see over here the permissions that are assigned to a role, or I can just create a role from scratch.

[00:46](javascript:;)So let me go to that.

[00:48](javascript:;)I'm going to click "Create role" and I'm going to give it a name.

[00:53](javascript:;)I'm going to call this the instance operator.

[00:57](javascript:;)There's also an ID to that and that must be unique and cannot be changed.

[01:03](javascript:;)There is a launch stage selection, Alpha, Beta, general availability, and disabled.

[01:10](javascript:;)This is essentially just a launch stage.

[01:13](javascript:;)So you want to make sure that you start small, tested Alpha, and then roll

[01:17](javascript:;)it out at some point so that other users know that they can leverage that availability.

[01:23](javascript:;)So what I'm going to do now is click "Add permissions" because currently there are no assigned permissions given that I started from scratch.

[01:30](javascript:;)So let's go in here.

[01:33](javascript:;)Now we have over 2,000 different permissions.

[01:36](javascript:;)So we obviously want to filter for that just a little bit and specifically I'm interested in the permissions for compute instances.

[01:43](javascript:;)So let me type compute.instances.

[01:50](javascript:;)and hit "Enter".

[01:53](javascript:;)Now I'm down to 44.

[01:56](javascript:;)So I want to select a couple different ones from here.

[01:59](javascript:;)I'm interested in "Get", I want to be able to get the different instances.

[02:04](javascript:;)I want to be able to list all of the instances as well as reset them and resume.

[02:11](javascript:;)Resume is if an instance was suspended which is equivalent to if it's sleep or in standby mode.

[02:17](javascript:;)I also want to start and stop, and suspend.

[02:25](javascript:;)So I can go click "Add" now and I can see the permissions that I just assigned.

[02:31](javascript:;)So I can get, list, reset, resume, start, stop, and suspend.

[02:35](javascript:;)From here, I can now click "Create".

[02:39](javascript:;)It's created that and I can click on it here.

[02:42](javascript:;)I can review that.

[02:44](javascript:;)I have an ID and I have a launch stage, and these are my permissions.

[02:51](javascript:;)That's how easy it is to create a custom role in GCP.

[02:55](javascript:;)Alternatively, I could have started with the instance admin role as a base and remove the permissions that I don't want the role to have.

[03:03](javascript:;)Now remember, that custom roles are not maintained by Google.

[03:07](javascript:;)That means that when new permissions, features or services are added to GCP, your custom roles will not be updated automatically.

Members-

[00:00](javascript:;)Let’s talk more about members, which define the “who” part of “who can do what on which resource.”

[00:06](javascript:;)There are five different types of members: Google Accounts, Service Accounts, Google Groups, Google Workspace domains, and Cloud Identity domains.

[00:16](javascript:;)A Google account represents a developer, an administrator, or any other person who interacts with Google Cloud.

[00:23](javascript:;)Any email address that is associated with a Google account can be an identity, including gmail.com or other domains.

[00:31](javascript:;)New users can sign up for a Google account by going to the Google account signup page, without receiving mail through Gmail.

[00:38](javascript:;)A service account is an account that belongs to your application instead of to an individual end user.

[00:44](javascript:;)When you run code that is hosted on Google Cloud, you specify the account that the code should run as.

[00:51](javascript:;)You can create as many service accounts as needed to represent the different logical components of your application.

[00:57](javascript:;)A Google group is a named collection of Google accounts and service accounts.

[01:02](javascript:;)Every group has a unique email address that is associated with the group.

[01:06](javascript:;)Google groups are a convenient way to apply an access policy to a collection of users.

[01:12](javascript:;)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.

[01:21](javascript:;)A Workspace domain represents a virtual group of all the Google accounts that have been created in an organization's Workspace account.

[01:29](javascript:;)Workspace domains represent your organization's internet domain name, such as example.com, and when you add a user to

[01:36](javascript:;)your Workspace domain, a new Google account is created for the user inside this virtual group, such as username@example.com.

[01:46](javascript:;)Google Cloud customers who are not Workspace customers can get these same capabilities through Cloud Identity.

[01:52](javascript:;)Cloud Identity lets you manage users and groups using the Google Admin Console, but you

[01:56](javascript:;)do not pay for or receive Workspace’s collaboration products such as Gmail, Docs, Drive, and Calendar.

[02:04](javascript:;)Now it’s important to note that you cannot use IAM to create or manage your users or groups.

[02:10](javascript:;)Instead, you can use Cloud Identity or Workspace to create and manage users.

[02:14](javascript:;)A policy consists of a list of bindings.

[02:16](javascript:;)A binding binds a list of members to a role, where the members can be user accounts, Google groups, Google domains, and service accounts.

[02:26](javascript:;)A role is a named list of permissions defined by IAM.

[02:30](javascript:;)Let’s revisit the IAM resource hierarchy.

[02:33](javascript:;)A policy is a collection of access statements attached to a resource.

[02:37](javascript:;)Each policy contains a set of roles and role members, with resources inheriting policies from their parent.

[02:44](javascript:;)Think of it like this: resource policies are a union of parent and resource, where a less restrictive parent policy will always override a more restrictive resource policy.

[02:55](javascript:;)The IAM policy hierarchy always follows the same path as the Google Cloud resource hierarchy, which means that if you change the resource hierarchy, the policy hierarchy also changes.

[03:08](javascript:;)For example, moving a project into a different organization will update the project's IAM policy to inherit from the new organization's IAM policy.

[03:18](javascript:;)Also, child policies cannot restrict access granted at the parent level.

[03:24](javascript:;)For example, if we grant you the Editor role for Department X, and we grant you

[03:28](javascript:;)the Viewer role at the bookshelf project level, you still have the Editor role for that project.

[03:34](javascript:;)Therefore, it is a best practice is to follow the principle of least privilege.

[03:40](javascript:;)The principle applies to identities, roles, and resources.

[03:44](javascript:;)Always select the smallest scope that’s necessary for the task in order to reduce your exposure to risk.

[03:50](javascript:;)You can also use a recommender for role recommendations to identify and remove excess permissions from your principals, improving your resources’ security configurations.

[04:00](javascript:;)Each role recommendation suggests that you remove or replace a role that gives your principals excess permissions.

[04:06](javascript:;)At scale, these recommendations help you enforce the principle of least privilege by ensuring that principals have only the permissions that they actually need.

[04:16](javascript:;)Recommender identifies excess permissions using policy insights.

[04:20](javascript:;)Policy insights are ML-based findings about permission usage in your project, folder, or organization.

[04:27](javascript:;)You can grant access to Google Cloud resources by using allow policies, also known as IAM policies, which are attached to resources.

[04:36](javascript:;)The allow policy controls access to the resource itself and any descendants of that resource that inherit the allow policy.

[04:44](javascript:;)An allow policy associates, or binds, one or more principals (also known as a member or identity)

[04:49](javascript:;)with a single IAM role and any context-specific conditions that change how and when the role is granted.

[04:57](javascript:;)In the example on this slide, Jie (jie@example.com) is granted the Organization Admin predefined role (roles/resourcemanager.organizationAdmin) in the first role binding.

[05:12](javascript:;)This role contains permissions for organizations, folders, and limited projects operations.

[05:18](javascript:;)In the second role binding, both Jie and Raha (raha@example.com) are granted the ability to create projects via the Project Creator role (roles/resourcemanager.projectCreator).

[05:31](javascript:;)Together, these role bindings grant fine-grained access to both Jie and Raha, and Jie is granted more access than Raha.

[05:43](javascript:;)IAM deny policies let you set guardrails on access to Google Cloud resources.

[05:49](javascript:;)With deny policies, you can define deny rules that prevent certain principals from using certain permissions, regardless of the roles they're granted.

[06:00](javascript:;)Deny policies are made up of deny rules.

[06:03](javascript:;)Each deny rule specifies a set of principals that are denied permissions, and the permissions that the principals are denied, or unable to use.

[06:14](javascript:;)Optionally, you can define the condition that must be true for the permission to be denied.

[06:20](javascript:;)When a principal is denied a permission, they can't do anything that requires that permission, regardless of the IAM roles they've been granted.

[06:30](javascript:;)This is because IAM always checks relevant deny policies before checking relevant allow policies.

[06:38](javascript:;)IAM Conditions allow you to define and enforce conditional, attribute-based access control for Google Cloud resources.

[06:45](javascript:;)With IAM Conditions, you can choose to grant resource access to identities (members) only if configured conditions are met.

[06:55](javascript:;)For example, this could be done to configure temporary access for users in the event of a

[07:00](javascript:;)production issue or to limit access to resources only for employees making requests from your corporate office.

[07:07](javascript:;)Conditions are specified in the role bindings of a resource's IAM policy.

[07:11](javascript:;)When a condition exists, the access request is only granted if the condition expression evaluates to true.

[07:18](javascript:;)Each condition expression is defined as a set of logic statements allowing you to specify one or more attributes to check.

[07:26](javascript:;)An organization policy is a configuration of restrictions, defined by configuring a constraint with the desired restrictions for that organization.

[07:35](javascript:;)An organization policy can be applied to the organization node, and all of its folders or projects within that node.

[07:44](javascript:;)Descendants of the targeted resource hierarchy inherit the organization policy that has been applied to their parents.

[07:52](javascript:;)Exceptions to these policies can be made, but only by a user who has the organization policy admin role.

[07:59](javascript:;)What if you already have a different corporate directory?

[08:02](javascript:;)How can you get your users and groups into Google Cloud?

[08:07](javascript:;)Using Google Cloud Directory Sync, your administrators can log in and manage Google Cloud resources using the same usernames and passwords they already use.

[08:17](javascript:;)This tool synchronizes users and groups from your existing Active Directory or LDAP system with the users and groups in your Cloud Identity domain.

[08:26](javascript:;)The synchronization is one-way only; which means that no information in your Active Directory or LDAP map is modified.

[08:33](javascript:;)Google Cloud Directory Sync is designed to run scheduled synchronizations without supervision, after its synchronization rules are set up.

[08:40](javascript:;)Google Cloud also provides single sign-on authentication.

[08:44](javascript:;)If you have your identity system, you can continue using your own system and processes with SSO configured.

[08:51](javascript:;)When user authentication is required, Google will redirect to your system.

[08:55](javascript:;)If the user is authenticated in your system, access to Google Cloud is given; otherwise, the user is prompted to sign in.

[09:03](javascript:;)This allows you to also revoke access to Google Cloud.

[09:06](javascript:;)If your existing authentication system supports SAML2, SSO configuration is as simple as 3 links and a certificate, as shown on this slide.

[09:16](javascript:;)Otherwise, you can use a third-party solution, like ADFS, Ping, or Okta.

[09:20](javascript:;)Also, if you want to use a Google account but are not interested in receiving mail through Gmail, you can still create an account without Gmail.

Service Accounts-

[00:00](javascript:;)As mentioned earlier, another type of member is a service account.

[00:03](javascript:;)A service account is an account that belongs to your application instead of to an individual end user.

[00:09](javascript:;)This provides an identity for carrying out server-to-server interactions in a project without supplying user credentials.

[00:16](javascript:;)For example, if you write an application that interacts with Google Cloud Storage, you must first authenticate to either Google Cloud Storage XML API or JSON API.

[00:27](javascript:;)You can enable service accounts and grant read-write access to the account on the instance where you plan to run your application.

[00:36](javascript:;)Then, program the application to obtain credentials from the service account.

[00:40](javascript:;)Your application authenticates seamlessly to the API without embedding any secret keys or credentials in your instance, image, or application code.

[00:49](javascript:;)Service accounts are identified by an email address like the example shown here.

[00:55](javascript:;)There are three types of service accounts: user-created or custom, built-in, and Google APIs service accounts.

[01:03](javascript:;)By default, all projects come with a built-in Compute Engine default service account.

[01:09](javascript:;)Apart from the default service account, all projects come with a Google Cloud APIs service account, identifiable by the email project-number@cloudservices.gserviceaccount.com.

[01:22](javascript:;)This is an account designed specifically to run internal Google processes on your behalf and is automatically granted the Editor role on the project.

[01:30](javascript:;)Alternatively, you can also start an instance with a custom service account.

[01:35](javascript:;)Custom service accounts provide more flexibility than the default service account, but they require more management from you.

[01:43](javascript:;)You can create as many custom service accounts as you need.

[01:46](javascript:;)Sign any arbitrary access scopes or Cloud IAM roles to them, and assigned the service accounts to any virtual machine instance.

[01:54](javascript:;)Let's talk more about the default Compute Engine service account.

[01:58](javascript:;)As I mentioned, this account is automatically created per project.

[02:03](javascript:;)This account is identifiable by the email project-number-compute@developer.gserviceaccount.com, and is automatically granted the Editor role on the project.

[02:15](javascript:;)When you start a new instance using gcloud, the default service account is embedded on that instance.

[02:23](javascript:;)You can override this behavior by specifying another service account or by disabling service accounts for the instance.

[02:30](javascript:;)Now ,authorization is the process of determining what permissions and authenticated identity has on a set of specified resources.

[02:39](javascript:;)Scopes are used to determine whether an authenticated identity is authorized.

[02:44](javascript:;)In the example shown here, applications A and B contain authenticated identities or service accounts.

[02:52](javascript:;)Let's assume that both applications want to use a Cloud Storage bucket.

[02:56](javascript:;)They each request access from the Google Authorization Server and in return they receive an access token.

[03:03](javascript:;)Application A receives an access token with read-only scope, so you can only read from the Cloud Storage bucket.

[03:10](javascript:;)Application B in contrast, receives an access token with read-write scope, so it can read and modify data in the Cloud

[03:16](javascript:;)Storage bucket, Scopes can be customized when you create an instance using the default service account as shown in the screenshot.

[03:25](javascript:;)These scopes can be changed after the instance is created by stopping it.

[03:30](javascript:;)Access scopes are actually a legacy method of specifying permissions for your VM.

[03:36](javascript:;)Before the existence of IAM roles, access scopes were the only mechanism for granting permissions to service accounts.

[03:43](javascript:;)For user-created service accounts use Cloud IAM roles instead to specify permissions.

[03:48](javascript:;)Now, roles for service accounts can also be assigned to groups or users.

[03:54](javascript:;)Let's look at the example shown on the slide.

[03:57](javascript:;)First, you create a service account that has the InstanceAdmin role, which has permissions to create, modify, and delete virtual machine instances and disks.

[04:08](javascript:;)Then, you treat the service account as a resource and decide who can use it by provisioning users or a group with the Service Account User role.

[04:19](javascript:;)This allows those users to act as that service account to create, modify, and delete virtual machine instances and disks.

[04:27](javascript:;)Users, who are Service Account Users for a service account can access all of the resources that service account has access to.

[04:35](javascript:;)Therefore, be cautious from granting the Service Account User role to a user or group.

[04:42](javascript:;)Here's another example.

[04:42](javascript:;)The VMS running component\_1 are granted Editor access to project\_b using Service Account

[04:49](javascript:;)1. The VMS running component\_2 are granted objectViewer access to bucket\_1 using an isolated Service Account

[04:57](javascript:;)2. In this way you can sculpt permissions for VMs without re-creating VMs, Essentially, Cloud IAM lets you slice

[05:05](javascript:;)up a project into different microservices, each with access to different resources, by creating service accounts to represent each one.

[05:14](javascript:;)You assign the service accounts to the VMs when they are created, and you don't

[05:18](javascript:;)need to ensure that credentials have been managed correctly because Google Cloud manages security for you.

[05:25](javascript:;)Now, you might ask, how are service accounts authenticated?

[05:32](javascript:;)There are two types of Google Service Accounts.

[05:34](javascript:;)By default, when using service accounts within Google Cloud, for example from Compute Engine or App Engine, Google automatically manages the keys for service accounts.

[05:44](javascript:;)However, if you want to be able to use service accounts outside of Google Cloud, or want a

[05:49](javascript:;)different rotation period, then it is possible to also manually create and manage your own service account keys.

[05:57](javascript:;)All service accounts have Google-managed key pairs.

[06:00](javascript:;)With Google-managed service account keys, Google stores both the public and private portion of the key, and rotates them regularly.

[06:09](javascript:;)Each public key can be used for signing for a maximum of two weeks.

[06:14](javascript:;)Your private key is always held securely in escrow and is never directly accessible.

[06:20](javascript:;)You may optionally create one or more user-managed key pairs, also known as "external" keys, that can be used from outside of Google Cloud.

[06:28](javascript:;)Google only stores the public portion of a user-managed key.

[06:32](javascript:;)The user is responsible for the security of the private key and performing other management operations such as key rotation, whether manually or programmatically.

[06:43](javascript:;)Users can create up to ten service account keys per service account to facilitate key rotation.

[06:49](javascript:;)User-managed keys can be managed by using the Cloud IAM API, the gcloud command-line tool, or the Service Account page in the Cloud Console.

[06:58](javascript:;)Google does not save your user-managed private keys, so if you lose them, Google cannot help you recover them.

[07:06](javascript:;)You are responsible for keeping these keys safe, and also responsible for performing key rotation.

[07:12](javascript:;)User-managed keys should be used as a last resort.

[07:16](javascript:;)Consider the other alternatives, such as short-lived service account credentials, (tokens) or service account impersonation.

[07:24](javascript:;)The gcloud command shown on this slide, is a fast and easy way to list all of the keys associated with a particular service account.

IAM Best Practices-

[00:00](javascript:;)Let's talk about some Cloud IAM best practices to help you apply the concepts you just learned in your day-to-day work.

[00:06](javascript:;)First, leverage and understand the resource hierarchy.

[00:11](javascript:;)Specifically, use projects to group resources that share the same trust boundary.

[00:16](javascript:;)Check the policy granted on each resource and make sure you recognize the inheritance.

[00:23](javascript:;)Because of inheritance, use the principle of least privilege when granting roles.

[00:28](javascript:;)Finally, audit policies using Cloud audit logs and audit memberships of groups using policies.

[00:35](javascript:;)Next, I recommend granting roles to groups instead of individuals.

[00:40](javascript:;)This allows you to update group membership instead of changing a Cloud IAM policy.

[00:45](javascript:;)If you do this, make sure to audit membership of groups used in policies and control the ownership of the Google group used in Cloud IAM policies.

[00:54](javascript:;)You can also use multiple groups to get better control.

[00:57](javascript:;)In the example on this slide, there is a network admin group.

[01:01](javascript:;)Some of those members also need a read write role to a Cloud Storage bucket, but others need the read only role.

[01:08](javascript:;)Adding and removing individuals from all three groups controls their total access.

[01:12](javascript:;)Therefore, groups are not only associated with job roles but can exist for the purpose of role assignment.

[01:18](javascript:;)Here are some best practices for using service accounts.

[01:22](javascript:;)As mentioned before, be very careful when granting the service accounts user role because it provides access to all the resources of the service account has access to.

[01:32](javascript:;)Also when you create a service account give it a display name that clearly identifies its purpose, ideally using an established naming convention.

[01:42](javascript:;)As for keys, establish key rotation policies and methods and audit keys with the serviceAccount.keys.list method.

[01:50](javascript:;)Finally, I recommend using Cloud Identity Aware Proxy or Cloud IAP.

[01:55](javascript:;)Cloud IAP lets you establish a central authorization layer for applications accessed by HTTPS.

[02:01](javascript:;)So you can use an application level access control model instead of relying on network level firewalls.

[02:07](javascript:;)Applications and resources protected by Cloud IAP can only be accessed through the proxy by users and groups with the correct Cloud IAM role.

[02:17](javascript:;)When you grant a user access to an application or resource by Cloud IAP, they're

[02:22](javascript:;)subject to the fine-grained access controls implemented by the product in use without requiring a VPN.

[02:29](javascript:;)Cloud IAP performs authentication and authorization checks when a user tries to access a Cloud IAP secure resource as shown on the right.

Lab Intro: Exploring IAM-

[00:00](javascript:;)It's time to apply what you learned.

[00:02](javascript:;)In this lab, you'll grant and revoke roles to change access.

[00:06](javascript:;)Specifically, you will use Cloud IAM to implement access control, restrict access to specific features and resources, and use the service account user role.

[00:16](javascript:;)Now, anytime you make changes to IAM roles, the GCP Console refreshes faster than the actual system.

[00:22](javascript:;)Therefore, you should expect some short delays when making changes to a member's role.

Exploring IAM- Lab to be done

Lab Review: Exploring IAM-

[00:00](javascript:;)In this lab, you granted and revoked Cloud IAM roles, first to a user username 2, and then to a service account user.

[00:08](javascript:;)Having access to both users allow you to see the results of the changes you made.

[00:12](javascript:;)You can stay for a lab walk-through, but remember that GCP's user interface can change.

[00:17](javascript:;)So your environment might look slightly different.

[00:19](javascript:;)Welcome to the walk-through of the Cloud IAM lab.

[00:24](javascript:;)In this lab, we have set up two users for you, and at this point I have logged into the console as username

[00:31](javascript:;)1. So Qwiklabs will provide you with two usernames to log into and we'll do some operations with both, but right now I have already logged in as username

[00:40](javascript:;)1. So the first instruction tells you to log into the console in another tab as username

[00:46](javascript:;)2. So console, I'm going to grab that username, certainly going here, I'm going to say add account, username 2,

[01:02](javascript:;)and luckily we've been given the same password for both usernames and login here, and with Qwiklabs you have new accounts.

[01:11](javascript:;)So it's always going to ask you for all of this new user acceptance, and I'll accept this terms, I'm good to go.

[01:23](javascript:;)So task two is to explore the IAM console.

[01:27](javascript:;)So I'm going to go to the username 1 tab.

[01:30](javascript:;)I'm going to go to IAM, and I'm going to click on there.

[01:35](javascript:;)If I hit Add, I can look around at the different roles I can provide.

[01:43](javascript:;)Let me go ahead and click Cancel, feel free to explore as much as you want,

[01:48](javascript:;)you can see here there are roles based on different products and services that we have.

[01:53](javascript:;)I'll hit Cancel.

[01:54](javascript:;)Let me go to username 2, and I'm going to do the same thing.

[01:59](javascript:;)Let me go to IAM.

[02:03](javascript:;)So I'm going to browse this list now and I am going to look for the names associated

[02:08](javascript:;)with username 1, which in my case ends in 82462, I see it here, and here is username

[02:17](javascript:;)2. You can see there are different roles associated with each one of them.

[02:21](javascript:;)Username 1 has App Engine admin, BigQuery admin, editor, owner, and viewer.

[02:27](javascript:;)Whereas, username 2, which is the one that I'm logged in in this tab only has Viewer Access.

[02:36](javascript:;)So now I'm going to move on to task three.

[02:40](javascript:;)So I'm going to go back to username 1, there's going to go to be a lot of switching back and forth in this lab.

[02:44](javascript:;)So make sure you keep track of which tab is username 1 and which is username

[02:48](javascript:;)2. I am going to go to Google Cloud Storage here, and I'm going to create a bucket in here.

[02:58](javascript:;)So buckets need to be globally unique.

[03:01](javascript:;)So I am going to use my Cloud Project ID because it is pretty unique, and I'm going to click Create and keep all

[03:11](javascript:;)the other defaults, and make sure to note the name of your bucket because we'll use it as your bucket name across the lab.

[03:24](javascript:;)So here I'm going to go to upload files, and let me find just any sample file here just a screenshot, and I've uploaded it there.

[03:41](javascript:;)Once it's uploaded, I'm going to rename it here, and I'm going to call it sample.txt.

[03:50](javascript:;)The reason I'm doing this is because it's going to be much easier to run any of the

[03:55](javascript:;)commands I'm going to do with sample.txt as a name as opposed to that long name I already had.

[04:01](javascript:;)So at this point in the lab, you can hit the Check my progress button inside the lab and it'll give you a green check and five points.

[04:09](javascript:;)If you've correctly created a bucket and uploaded a sample file.

[04:14](javascript:;)So now I'm going to switch to username 2 and I'm going to go to storage browser.

[04:24](javascript:;)I'm going to verify that username 2 has view access to that bucket, and here it is.

[04:30](javascript:;)Because it's inherited that, I can view the sample file.

[04:34](javascript:;)Task four is I'm going to remove project viewer role for username

[04:39](javascript:;)2. So in order to do that I have to go back to username

[04:42](javascript:;)1. I'm going to go back to IAM, and then I'm going to find username 2 which is this one right here, 73.

[04:53](javascript:;)I'm going to hit Edit, and then I am going to hit the garbage can so that I can remove it.

[05:01](javascript:;)I hit Save, and then I at this point I can also check my progress and I'll get five points and a green check mark.

[05:13](javascript:;)I should have 10 points out of 20 in the lab.

[05:15](javascript:;)If I have properly done that.

[05:18](javascript:;)Throughout these labs if you ever get to a point where you realize that you didn't get the points necessary, it's probably

[05:24](javascript:;)because you missed a step or two, granted sometimes the lab is actually broken because they're based on technology that changes a lot.

[05:33](javascript:;)But if a lab isn't broken, chances are you just missed a step.

[05:38](javascript:;)So I usually recommend go back three steps.

[05:41](javascript:;)Check your work, make sure you did everything.

[05:44](javascript:;)Usually, that's what happened.

[05:46](javascript:;)So now we're going to verify that the username 2 has lost access.

[05:51](javascript:;)So I'm going to go back to the username 2 bucket tab, and then I'm going to click Home, and then I'm going to go back to storage to verify.

[06:03](javascript:;)I could've just refresh the screen as well, Refresh.

[06:08](javascript:;)List of buckets could not be loaded.

[06:11](javascript:;)So as you can see I do not have access anymore.

[06:15](javascript:;)So now the next task, task five, is to add Storage Access.

[06:19](javascript:;)So I'm going to copy the value of username 2 from the Qwiklabs lab name, from their connection details on the left of your lab instructions.

[06:30](javascript:;)So I'll copy that, I'm going to go back to username 1 tab, and I'm already in I am.

[06:36](javascript:;)So I'm going to hit Add, and then for new members I'm going to paste the value here.

[06:44](javascript:;)That is it, and I am going to select Storage, scroll down.

[06:51](javascript:;)Luckily it's alphabetical, and I am giving it Storage Object Viewer.

[06:58](javascript:;)Then I'm going to hit Save.

[07:01](javascript:;)This is another checkpoint in the lab where you can go back and hit Check my progress, and

[07:05](javascript:;)you should get another five points that you have checked, that you have actually provided the right permissions.

[07:12](javascript:;)Now we had one in the modules that sometimes the permissions upload faster than will be displayed in the GCP Console.

[07:19](javascript:;)Sometimes you just have to be patient.

[07:22](javascript:;)Maybe click Check my progress, wait a couple seconds if you didn't get it, and then you'll get the five points and the green check.

[07:28](javascript:;)So the next piece of task 5 is to verify that username 2 now has storage access.

[07:34](javascript:;)So if I go back here, and I'm going to start Cloud Shell.

[07:44](javascript:;)Because username 2 doesn't have project viewer roles, so it won't be able to see anything in the console, but we can see things in Cloud Storage.

[07:54](javascript:;)So we're going to use Cloud Shell for that.

[07:56](javascript:;)So let's make sure I know my bucket name.

[08:00](javascript:;)I copied it earlier, but I have definitely forgotten it by now.

[08:04](javascript:;)So let me go back here to Storage easily, and I can easily copy paste the bucket name here.

[08:12](javascript:;)Copy that, and back in Cloud Shell, I am going to do a gsutil ls to list for that bucket,

[08:23](javascript:;)gs://my bucket name, and username 2 should be able to see that there's sample.txt in the bucket and there it is.

[08:32](javascript:;)So now I can close username 2 tab because the rest of the lab is done in the username 1 console.

[08:42](javascript:;)So task 6 is to set up the service account user.

[08:46](javascript:;)So in IAM, I'm going to go to Service accounts.

[08:53](javascript:;)I'm going to create a service account, and the service account name is going to be read-bucket-objects, and I'm going to hit Create.

[09:05](javascript:;)It's going to ask me which role to provide, and I am going to be giving it Storage, Storage Object Viewer.

[09:13](javascript:;)Hit Continue, and then I'm going to hit Done.

[09:19](javascript:;)So now we've created our service account, so we're going to go back to the main IAM page,

[09:25](javascript:;)and we are going to select the service account we just created, and we're going to hit Add.

[09:33](javascript:;)In order to add members, normally you could perform this activity for a specified user group or a domain.

[09:42](javascript:;)But for training purposes and for this video, we're just going to grant the service account user

[09:45](javascript:;)role to everyone at a company called autostrat.com, which is a fake company used for demonstrating and training.

[09:53](javascript:;)So the new member is going to be autostrat.com, and I am going to give it Service Accounts, Service Account User, and then I'm going to hit Save.

[10:15](javascript:;)So now I'm going to go back to IAM, and I am going to add, and I am going to provide compute engine access.

[10:33](javascript:;)So the new member is autostrat.com.

[10:39](javascript:;)Make sure you're typing it correctly.

[10:41](javascript:;)I am giving it Compute Engine, and Compute Instance Admin V1 and save.

[10:56](javascript:;)So essentially, that step is a rehearsal of activity that you would probably perform for a specific user.

[11:02](javascript:;)It gives the user limited abilities with a VM Instance.

[11:06](javascript:;)It would be able to connect via SSH to a VM and perform possibly some administration tasks.

[11:12](javascript:;)So now, I am going to create a VM with the service account I created.

[11:23](javascript:;)Create, I am going to use the same name provided in a lab, demoIAM.

[11:31](javascript:;)I'm using us-central1.

[11:33](javascript:;)The zone is us-central1-c, and the machine type is an F1-micro.

[11:44](javascript:;)It is just for demonstration purposes.

[11:46](javascript:;)So let's not waste resources, and the service account is the read bucket objects account, and I'm going to hit Create.

[11:57](javascript:;)So this is another checkpoint in the lab, and you should be able to hit

[12:01](javascript:;)Check my progress, and verify that you have gotten the last five points in the lab.

[12:07](javascript:;)Again, this is another one that might take a couple seconds to propagate.

[12:12](javascript:;)So just give it a second, and make sure that you get the green check in the final five points.

[12:20](javascript:;)Task 7, you explore the service account user role, and now you've already completed all of the tasks in the labs.

[12:27](javascript:;)So this is just for learning purposes.

[12:30](javascript:;)So I'm going to go in here, and I'm going to SSH into this account into the VM that I just created.

[12:37](javascript:;)Then I am going to run a gcloud compute instances list.

[12:45](javascript:;)I am expecting to see an error because I do not have the correct permissions to list those for my project.

[12:58](javascript:;)Just wait for that to show up, and there you can see error.

[13:03](javascript:;)Some request did not succeed because I don't have permission to do that.

[13:06](javascript:;)So now I'm going to try to copy the file from the bucket that I created earlier.

[13:15](javascript:;)So my bucket name is the Project ID which I've forgotten already.

[13:25](javascript:;)Here gets/sample.txt, and you can see it successfully copied it.

[13:38](javascript:;)Now I'm going to copy it into another file, and then I'm going to try to upload into my bucket.

[13:52](javascript:;)Bucket name is here, and you'll see I can download, but I cannot add.

[14:08](javascript:;)In review in this lab, you granted and revoked Cloud IAM roles, first for user, and then to service account user.

[14:17](javascript:;)I hope you enjoyed the walkthrough.

[14:19](javascript:;)Thank you.

Quiz: Identity and Access Management-

Which of the following is not a type of IAM member?

Organization Account

That's correct! There are five different types of members: Google Accounts, Service Accounts, Google Groups, Google Workspace domain, and Cloud Identity domains. There are no "Organization Accounts" in IAM.

Which of the following is not a type of IAM role?

Advanced

That's correct! There are three types of roles in IAM: basic roles, predefined roles, and custom roles. There are no "advanced" roles in IAM.

What abstraction is primarily used to administer user access in IAM ?

Roles, an abstraction of job roles.

IAM administration uses pre-defined roles for administration of user access. The roles are defined by more granular permissions. But permissions are not applied to users directly, only through the roles that are assigned to them.

Module Review-

[00:00](javascript:;)Person: In this module, we covered Identity and Access Management along with its components and best practices.

[00:06](javascript:;)IAM builds on top of other Google Cloud and entity services.

[00:11](javascript:;)The creation and administration of corporate identities occurs through the workspace admin or cloud identity interface and is commonly handled by a person separate from the Google Cloud administrator.

[00:24](javascript:;)Google Groups are a great way for these two business functions to collaborate.

[00:29](javascript:;)You establish the roles and assign them through the group, and then the workspace admin administers membership in the group.

[00:37](javascript:;)Finally, remember that service accounts are very flexible, and that they can enable you to build an infrastructure-based level of control in your application.

STORAGE AND DATABASE SERVICES:

Module Overview-

[00:00](javascript:;)In this module, we cover storage and database services in Google Cloud.

[00:05](javascript:;)Every application needs to store data, whether it's business data, media to be streamed, or sensor data from devices.

[00:13](javascript:;)From an application-centered perspective, the technology stores and retrieves the data.

[00:19](javascript:;)Whether it's a database or an object store is less important than whether that service supports the application’s requirements for efficiently storing and retrieving the data, given its characteristics.

[00:31](javascript:;)Google offers several data storage services to choose from.

[00:36](javascript:;)In this module, we will cover Cloud Storage, Filestore, Cloud SQL, Cloud Spanner, Cloud Firestore, and Cloud Bigtable.

[00:46](javascript:;)Let me start by giving you a high-level overview of these different services.

[00:51](javascript:;)This table shows the storage and database services and highlights the storage service type (object, file, relational, non-relational, and data warehouse), what each service is good for, and intended use.

[01:05](javascript:;)BigQuery is also listed on the right.

[01:08](javascript:;)I’m mentioning this service because it sits on the edge between data storage and data processing.

[01:13](javascript:;)You can store data in BigQuery, but the intended use for BigQuery is big data analysis and interactive querying.

[01:21](javascript:;)For this reason, BigQuery is covered later in the course.

[01:26](javascript:;)If tables aren’t your preference, here’s a decision tree to help you identify the solution that best fits your application.

[01:33](javascript:;)Let’s walk through this together.

[01:35](javascript:;)First, ask yourself: Is your data structured?

[01:40](javascript:;)If its not, then ask yourself if you need a shared file system.

[01:46](javascript:;)If you do, then choose Filestore.

[01:48](javascript:;)If you don't, then choose Cloud Storage.

[01:51](javascript:;)If your data is structured, does your workload focus on analytics?

[01:55](javascript:;)If your data is structured, does your workload focus on analytics?

[01:59](javascript:;)If it does, you will want to choose Bigtable or BigQuery, depending on your latency and update needs.

[02:07](javascript:;)BigQuery is recommended as a data warehouse, is the default storage for tabular data, and is optimized for large-scale, ad-hoc SQL-based analysis and reporting.

[02:19](javascript:;)While BigQuery data manipulation language (DML) enables you to update, insert, and delete data from your BigQuery tables,

[02:27](javascript:;)because it has a built-in cache BigQuery works really well in cases where the data does not change often.

[02:36](javascript:;)Bigtable is a NoSQL wide-column database.

[02:40](javascript:;)It's optimized for low latency, large numbers of reads and writes, and maintaining performance at scale.

[02:47](javascript:;)In addition to analytics, Bigtable is also suited as a ‘fast lookup’ non-relational database for datasets

[02:53](javascript:;)too large to store in memory, with use cases in areas such as IoT, AdTech and FinTec.

[03:02](javascript:;)If your workload doesn’t involve analytics, check whether your data is relational.

[03:08](javascript:;)If it’s not relational, do you need application caching?

[03:12](javascript:;)If caching is a requirement, choose Memorystore, an in-memory database.

[03:17](javascript:;)Otherwise choose Firestore, a document database.

[03:21](javascript:;)If your data is relational, you will want to choose Cloud SQL or Cloud Spanner, depending on your need for horizontal scalability.

[03:29](javascript:;)If you need horizontal scaling and a globally available system, Cloud Spanner is a good choice.

[03:35](javascript:;)Otherwise, Cloud SQL is a cost-effective solution.

[03:39](javascript:;)Depending on your application, you might use one or several of these services to get the job done.

[03:45](javascript:;)For more information on how to choose between these different services, please refer to the links provided in the course resources for this module.

[03:53](javascript:;)Before we dive into each of the data storage services, let’s define the scope of this module.

[04:00](javascript:;)The purpose of this module is to explain which services are available and when to consider using them from an infrastructure perspective.

[04:06](javascript:;)I want you to be able to set up and connect to a service without detailed knowledge of how to use a database system.

[04:15](javascript:;)If you want a deeper dive into the design, organizations, structures, schemas and details on how data

[04:20](javascript:;)can be optimized, served and stored properly within those different services, I recommend Google Cloud’s Data Engineering courses.

[04:29](javascript:;)Let’s look at the agenda.

[04:31](javascript:;)This module covers all of the services we have mentioned so far.

[04:35](javascript:;)To become more comfortable with these services, you will apply them in two labs.

[04:39](javascript:;)I’ll also provide a quick overview of Memorystore, which is Google Cloud’s fully managed Redis service.

[04:47](javascript:;)Let’s get started by diving into Cloud Storage and Filestore!

Cloud Storage-

[00:00](javascript:;)Cloud Storage is Google Cloud’s object storage service, and it allows world-wide storage and retrieval of any amount of data at any time.

[00:08](javascript:;)You can use Cloud Storage for a range of scenarios including serving website content, storing

[00:12](javascript:;)data for archival and disaster recovery, or distributing large data objects to users via direct download.

[00:20](javascript:;)Cloud Storage has a couple of key features: It’s scalable to exabytes of data; The time to first byte is

[00:26](javascript:;)in milliseconds; It has very high availability across all storage classes; And It has a single API across those storage classes.

[00:37](javascript:;)Some like to think of Cloud Storage as files in a file system but it’s not really a file system.

[00:42](javascript:;)Instead, Cloud Storage is a collection of buckets that you place objects into.

[00:47](javascript:;)You can create directories, so to speak, but really a directory is just another object that points to different objects in the bucket.

[00:54](javascript:;)You’re not going to easily be able to index all of these files like you would in a file system.

[01:00](javascript:;)You just have a specific URL to access objects.

[01:04](javascript:;)Cloud Storage has four storage classes: Standard, Nearline, Coldline and Archive, and each of those storage classes provide 3 location

[01:12](javascript:;)types: There’s a multi-region is a large geographic area, such as the United States, that contains two or more geographic places.

[01:23](javascript:;)Dual-region is a specific pair of regions, such as Finland and the Netherlands.

[01:28](javascript:;)A region is a specific geographic place, such as London.

[01:33](javascript:;)Objects stored in a multi-region or dual-region are geo-redundant.

[01:37](javascript:;)Now, let’s go over each of the storage classes: Standard Storage is best for data

[01:42](javascript:;)that is frequently accessed (think of "hot" data) and/or stored for only brief periods of time.

[01:50](javascript:;)This is the most expensive storage class but it has no minimum storage duration and no retrieval cost.

[01:57](javascript:;)When used in a region, Standard Storage is appropriate for storing data in the same location as Google Kubernetes Engine clusters or Compute Engine instances that use the data.

[02:08](javascript:;)Co-locating your resources maximizes the performance for data-intensive computations and can reduce network charges.

[02:15](javascript:;)When used in a dual-region, you still get optimized performance when accessing Google Cloud products that are located in

[02:20](javascript:;)one of the associated regions, but you also get improved availability that comes from storing data in geographically separate locations.

[02:30](javascript:;)When used in multi-region, Standard Storage is appropriate for storing data that is accessed around the world,

[02:36](javascript:;)such as serving website content, streaming videos, executing interactive workloads, or serving data supporting mobile and gaming applications.

[02:46](javascript:;)Nearline Storage is a low-cost, highly durable storage service for storing infrequently accessed data like data backup, long-tail multimedia content, and data archiving.

[02:57](javascript:;)Nearline Storage is a better choice than Standard Storage in scenarios where slightly lower availability, a

[03:02](javascript:;)30-day minimum storage duration, and costs for data access are acceptable trade-offs for lowered at-rest storage costs.

[03:12](javascript:;)Coldline Storage is a very-low-cost, highly durable storage service for storing infrequently accessed data.

[03:19](javascript:;)Coldline Storage is a better choice than Standard Storage or Nearline Storage in scenarios where slightly lower availability,

[03:24](javascript:;)a 90-day minimum storage duration, and higher costs for data access are acceptable trade-offs for lowered at-rest storage costs.

[03:36](javascript:;)Archive Storage is the lowest-cost, highly durable storage service for data archiving, online backup, and disaster recovery.

[03:43](javascript:;)Unlike the so-to-speak "coldest" storage services offered by other Cloud providers, your data is available within milliseconds, not hours or days.

[03:52](javascript:;)Unlike other Cloud Storage classes, Archive Storage has no availability SLA, though the typical availability is comparable to Nearline and Coldline Storage.

[04:03](javascript:;)Archive Storage also has higher costs for data access and operations, as well as a 365-day minimum storage duration.

[04:12](javascript:;)Archive Storage is the best choice for data that you plan to access less than once a year.

[04:17](javascript:;)Let’s focus on durability and availability.

[04:20](javascript:;)All of these storage classes have 11 nines of durability, but what does that mean?

[04:26](javascript:;)Does that mean you have access to your files at all times?

[04:29](javascript:;)No, what that means is you won't lose data.

[04:32](javascript:;)You may not be able to access the data, which is like going to your bank and saying well my money is in there, it's 11 nines durable.

[04:39](javascript:;)But when the bank is closed we don't have access to it, which is the availability that differs between storage classes and the location type.

[04:47](javascript:;)Cloud Storage is broken down into a couple of different items here.

[04:51](javascript:;)First of all, there are buckets which are required to have a globally unique name and cannot be nested.

[04:59](javascript:;)The data that you put into those buckets are objects that inherit the storage class of the bucket and those objects could be text files, doc files, video files, etc.

[05:11](javascript:;)There is no minimum size to those objects and you can scale this as much as you want as long as your quota allows it.

[05:19](javascript:;)To access the data, you can use the gcloud storage command, or either the JSON or XML APIs.

[05:29](javascript:;)When you upload an object to a bucket, the object is assigned the bucket's storage class, unless you specify a storage class for the object.

[05:38](javascript:;)You can change the default storage class of a bucket but you can't change the location type from regional to multi-region/dual-region or vice versa.

[05:47](javascript:;)You can also change the storage class of an object that already exists in your

[05:50](javascript:;)bucket without moving the object to a different bucket or changing the URL to the object.

[05:56](javascript:;)Setting a per-object storage class is useful, for example, if you have objects in your bucket that you want to keep, but that you don't expect to access frequently.

[06:06](javascript:;)In this case, you can minimize costs by changing the storage class of those specific objects to Nearline, Coldline or Archive Storage.

[06:13](javascript:;)In order to help manage the classes of objects in your bucket, Cloud Storage offers Object Lifecycle Management.

[06:20](javascript:;)More on that later.

[06:21](javascript:;)Let’s look at access control for your objects and buckets that are part of a project.

[06:27](javascript:;)We can use IAM for the project to control which individual user or service account can see the bucket,

[06:32](javascript:;)list the objects in the bucket, view the names of the objects in the bucket, or create new buckets.

[06:39](javascript:;)For most purposes, IAM is sufficient, and roles are inherited from project to bucket to object.

[06:46](javascript:;)Access control lists or ACLs offer finer control.

[06:50](javascript:;)For even more detailed control, signed URLs provide a cryptographic key that gives time-limited access to a bucket or object.

[06:58](javascript:;)Finally, a signed policy document further refines the control by determining what kind of file can be uploaded by someone with a signed URL.

[07:08](javascript:;)Let’s take a closer look at ACLs and signed URLs.

[07:12](javascript:;)An ACL is a mechanism you use to define who has access to your buckets and objects, as well as what the level of access is they have.

[07:21](javascript:;)The maximum number of ACL entries you can create for a bucket or object is 100.

[07:26](javascript:;)Each ACL consists of one or more entries, and these entries consist of two pieces of information: A

[07:33](javascript:;)scope, which defines who can perform the specified actions (for example, a specific user or group of users).

[07:41](javascript:;)And a permission, which defines what actions can be performed (for example, read or write).

[07:47](javascript:;)The allUsers identifier listed on this slide represents anyone who is on the internet, with or without a Google account.

[07:55](javascript:;)The allAuthenticatedUsers identifier, in contrast, represents anyone who is authenticated with a Google account.

[08:03](javascript:;)For more information on ACLs, refer to the links of this video.

[08:09](javascript:;)For some applications, it is easier and more efficient to grant limited-time access tokens that can be used by any user, instead of using account-based authentication for controlling resource access.

[08:20](javascript:;)(For example, when you don’t want to require users to have a Google account).

[08:25](javascript:;)Signed URLs allow you to do this for Cloud Storage.

[08:28](javascript:;)You create a URL that grants read or write access to a specific Cloud Storage resource and specifies when the access expires.

[08:36](javascript:;)That URL is signed using a private key associated with a service account.

[08:41](javascript:;)When the request is received, Cloud Storage can verify that the access-granting URL was issued on behalf of a trusted

[08:48](javascript:;)security principal, in this case the service account, and delegates its trust of that account to the holder of the URL.

[08:57](javascript:;)After you give out the signed URL, it is out of your control.

[09:00](javascript:;)So you want the signed URL to expire after some reasonable amount of time.

Cloud Storage Features-

[00:00](javascript:;)There are also several features that come with Cloud Storage.

[00:04](javascript:;)I will cover these at a high level for now because we will soon dive deeper into some of them.

[00:09](javascript:;)Earlier in the core series, we already talked a little bit about customer-supplied encryption keys when attaching persistent disks to virtual machines.

[00:18](javascript:;)This allows you to supply your own encryption keys instead of the Google-managed keys, which is also available for Cloud Storage.

[00:25](javascript:;)Cloud Search also provides Object Lifecycle Management, which lets you automatically delete or archive objects.

[00:32](javascript:;)Another feature is Object Versioning, which allows you to maintain multiple versions of objects in your pocket.

[00:38](javascript:;)You are charged for the versions as if there are multiple files, which is something to keep in mind.

[00:44](javascript:;)Cloud Storage also offers directory synchronization so that you can sync a VM directory with a bucket.

[00:51](javascript:;)We will discuss object change notification, data import, and strong consistency in more detail after going into Object Versioning and Object Lifecycle Management.

[01:01](javascript:;)In Cloud Storage, objects are immutable, which means that an uploaded object cannot change throughout its storage lifetime.

[01:08](javascript:;)To support the retrieval of objects that are deleted or are written, Cloud Storage offers the Object Versioning feature.

[01:15](javascript:;)Object Versioning can be enabled for a bucket.

[01:18](javascript:;)Once enabled, Cloud Storage creates an archived version of an object each time the live version of the object is overwritten or deleted.

[01:26](javascript:;)The archive version retains the name of the object, but is uniquely identified by a generation number as illustrated on this slide by g1.

[01:35](javascript:;)When Object Versioning is enabled, you can list archived versions of an object, restore the live

[01:40](javascript:;)version of an object to an older state, or permanently delete an archived version as needed.

[01:46](javascript:;)You can turn versioning on or off or a bucket at anytime.

[01:50](javascript:;)Turning versioning off leaves existing object versions in place and causes the bucket to stop accumulating new archived object versions.

[02:00](javascript:;)For more information on Object Versioning, refer to the link section of this video.

[02:05](javascript:;)To support common use cases like setting a time to live for objects, archiving older versions of

[02:11](javascript:;)objects, or downgrading storage classes of objects to help manage costs, Cloud Storage offers Object Lifecycle Management.

[02:19](javascript:;)You can assign a lifecycle management configuration to a bucket.

[02:23](javascript:;)The configuration is a set of rules that apply to all objects in the buckets.

[02:29](javascript:;)When an object meets the criteria of one of the rules, Cloud Storage automatically performs a specified action on the object.

[02:37](javascript:;)Here are some example use cases.

[02:39](javascript:;)First, downgrade the storage class of objects older than a year to Cloud line storage.

[02:45](javascript:;)Second, delete objects created before a specific date, for example, January 1st, 2017.

[02:51](javascript:;)Third, keep only the three most recent versions of each object, any bucket with versioning-enabled.

[02:58](javascript:;)Object inspection occurs in asynchronous batches, so rules may not be applied immediately.

[03:04](javascript:;)Also, updates to your lifecycle configuration may take up to 24-hours to go into effect.

[03:10](javascript:;)This means that when you change your lifecycle configuration, Object Lifecycle Management may still perform actions based on the old configuration for up to 24-hours, so keep that in mind.

[03:22](javascript:;)For more on Object Lifecycle Management, refer to the link section of this video.

[03:28](javascript:;)The Cloud Console, allows you to upload individual files to a bucket.

[03:33](javascript:;)But what if you have to upload terabytes, even petabytes of data?

[03:36](javascript:;)There are three services that address this: Transfer Appliance, Storage Transfer Service, and Offline Media Import.

[03:44](javascript:;)Transfer Appliance is a hardware appliance you can use to securely migrate large volumes of

[03:49](javascript:;)data from hundreds of terabytes up to one petabyte to Google Cloud without disrupting business operations.

[03:56](javascript:;)The images on the slide are Transfer Appliances.

[04:00](javascript:;)The Storage Transfer Service enables high-performance imports of online data.

[04:05](javascript:;)The data source can be another Cloud Storage bucket, an Amazon S3 bucket, or an HTTP, HTTPS location.

[04:13](javascript:;)Finally, offline media import is a third-party service, where physical media such as storage arrays, hard

[04:19](javascript:;)disk drives, tapes, and USB flash drives is sent to a provider who uploads the data.

[04:26](javascript:;)For more information on these three services, refer to the links section of this video.

[04:32](javascript:;)When you upload an object to Cloud Storage and you receive a success response, the

[04:37](javascript:;)object is immediately available for download and metadata operations from any location where Google offers service.

[04:45](javascript:;)This is true whether you create a new object or overwrite an existing object.

[04:51](javascript:;)Because uploads are strongly consistent, you will never receive a 404 NOT Found response or stale data for a read-after-write or read-after-metadata-update operation.

[05:04](javascript:;)Strong global consistency also extends to the deletion operation of objects.

[05:10](javascript:;)If a deletion request succeeds, an immediate attempt to download the object or its metadata will result in a 404 Not Found status code.

[05:18](javascript:;)You get the 404 error because the object no longer exists after the delete operation succeeds.

[05:24](javascript:;)Bucket listing is strongly consistent.

[05:28](javascript:;)For example, if you create a bucket, then immediately perform a list bucket operation, the new bucket appears in the returned list of buckets.

[05:36](javascript:;)Finally, object listing is also strongly consistent.

[05:40](javascript:;)For example, if you upload an object to a bucket and then immediately perform a list object operations, the new object appears in the returned list of objects.

Choosing a Storage Class-

[00:00](javascript:;)Let's explore the decision tree to help you find the appropriate storage class in Cloud storage.

[00:05](javascript:;)If you will read your data less than once a year, you should consider using Archive Storage.

[00:11](javascript:;)If you will read your data less than once per 90 days, you should consider using Coldline Storage.

[00:17](javascript:;)If you read your data less than once per 30 days, you should consider using Nearline Storage.

[00:23](javascript:;)If you'll be doing reads and writes more often that, you should consider using Standard Storage.

[00:29](javascript:;)You also want to take into account the location type, use a region to help optimize latency

[00:35](javascript:;)and network bandwidth for data consumers, such as analytics pipelines that are grouped in the same region.

[00:42](javascript:;)Use a dual-region when you want similar performance advantages as regions.

[00:46](javascript:;)But you also want the high availability that comes with being geo-redundant.

[00:50](javascript:;)Use a multi-region when you want to serve content to data consumers that are outside of the Google network,

[00:57](javascript:;)and distributed across large geographic areas, or when you want the higher data availability that comes with being geo-redundant.

Filestore-

[00:00](javascript:;)Filestore is a managed file storage service for applications that require a file system interface and a shared file system for data.

[00:08](javascript:;)Filestore gives users a simple native experience for standing up managed network attached storage with either Compute Engine or Google Kubernetes Engine instances.

[00:19](javascript:;)The ability to fine tune Filestore's performance and capacity independently, leads to predictably fast performance for your file-based workloads.

[00:28](javascript:;)Filestore offers native compatibility with existing enterprise applications and supports any NFSV3 compatible clients.

[00:37](javascript:;)Applications gain the benefit of features such as scale-out performance, hundreds of terabytes of capacity,

[00:43](javascript:;)and file locking without the need to install or maintain any specialized plug-ins or client-side software.

[00:50](javascript:;)Filestore has many use cases.

[00:52](javascript:;)Using Filestore, you can expedite migration of enterprise applications.

[00:57](javascript:;)Many on-premises applications require a file system interface to data.

[01:01](javascript:;)As these applications continue to migrate to the Cloud, Filestore can support a broad range of enterprise applications that need a shared file system.

[01:10](javascript:;)For media rendering, you can easily meant filestore file shares on Compute Engine instances, enabling visual effects artists to collaborate on the same file share.

[01:20](javascript:;)As rendering workflows typically run across fleets of Compute Machines, all of which meant a shared file system.

[01:28](javascript:;)Filestore and Compute Engine can scale to meet your jobs rendering needs.

[01:33](javascript:;)Electronic Design Automation, or EDA, is all about data management.

[01:38](javascript:;)It requires the ability to batch workloads across thousands of cores and has a large memory needs.

[01:44](javascript:;)Filestore offers the necessary capacity and scale to meet the needs of manufacturing customers doing intensive EDA, and also make sure that files are universally accessible.

[01:56](javascript:;)Data analytics workloads include Compute complex financial models or analysis of environmental data.

[02:03](javascript:;)These workloads are latency sensitive.

[02:06](javascript:;)Filestore offers low latency for file operations, and as capacity or performance needs change, you can easily grow or shrink your instances as needed.

[02:17](javascript:;)As a persistent and shareable storage layer, Filestore enables immediate access to data for high-performance, smart

[02:23](javascript:;)analytics without the need to lose valuable time on loading an offloading data to clients drives.

[02:31](javascript:;)Genome sequencing requires an incredible amount of raw data in the order of billions of data points per person.

[02:38](javascript:;)This type of analysis requires speed, scalability, and security.

[02:43](javascript:;)Filestore meets the needs of companies and research institutions performing scientific research while also offering predictable prices for the performance.

[02:52](javascript:;)Web developers and large hosting providers also rely on Filestore to manage and serve web content, including needs such as WordPress hosting.

Lab Intro: Cloud Storage-

[00:00](javascript:;)Let's take some of the cloud storage concepts that we just discussed and apply them in a lab.

[00:05](javascript:;)In this lab, you'll create buckets and perform many of the advanced options available in cloud storage.

[00:10](javascript:;)You'll set access control list to limit who can have access to your data and what they're allowed to do with it.

[00:16](javascript:;)You'll use the ability to supply and manage your own encryption keys for additional security.

[00:22](javascript:;)You'll enable object versioning to track changes in the data and you'll configure lifecycle management, so that objects are automatically archived or deleted after a specified period.

[00:33](javascript:;)Finally, you'll use the directory synchronization feature that I mentioned and share your buckets across projects using Cloud IAM.

Cloud Storage- Lab to be done

Lab Review: Cloud Storage-

[00:00](javascript:;)In this lab, you learn to create and work with Buckets and Objects and apply the following Cloud Storage

[00:04](javascript:;)features, Customer Supplied Encryption Keys, Access Control Lists, Life-cycle Management, Object Versioning, Directory Synchronization, and Cross-Project Resource Sharing using IAM.

[00:18](javascript:;)Now that you're familiar with many of the advanced features of Cloud storage, you might consider using them in a variety of applications that you might not have previously considered.

[00:27](javascript:;)A common, quick, and easy way to start using GCP is to use Cloud storage as a backup service.

[00:33](javascript:;)You can stay for a lab walk-through, but remember that GCP's user interface can change.

[00:38](javascript:;)So your environment might look slightly different.

[00:41](javascript:;)Welcome to the walk-through of the Cloud Storage Lab.

[00:45](javascript:;)At this point, I've already started the lab in Qwiklabs and I am logged into the GCP console

[00:50](javascript:;)using the username and password that was provided by Qwiklabs for me to log in to the GCP console.

[00:56](javascript:;)So the first task is preparation.

[00:58](javascript:;)I'm going to create a bucket in here.

[01:03](javascript:;)When I go to create a bucket, it specifically tells me that I should be using a globally unique ID.

[01:09](javascript:;)So I'm going to use my project ID, which is pretty unique.

[01:13](javascript:;)I'm going to call it myproj- and then my project ID, and it's telling us multi-regional.

[01:21](javascript:;)So storage class is multi-regional, and then it's telling me access control is set object-level and bucket-level permissions, and I'm going to hit create.

[01:36](javascript:;)So at this point, you can now go back to the lab page, and you can hit check

[01:43](javascript:;)my progress, and you should get a check mark in five points that you created the Cloud Storage Bucket.

[01:50](javascript:;)Next step is downloading a file.

[01:54](javascript:;)So I'm going to start Cloud Shell so I can do the curl command, and the first thing I'm going to do is I'm

[02:01](javascript:;)going to set an environment variable to the bucket name of the bucket I just created, just for ease of copy paste of commands.

[02:09](javascript:;)Export bucket name one equals and the bucket name.

[02:16](javascript:;)If I want to verify that that worked, I'm going to do an echo dollar sign,

[02:21](javascript:;)and the variable name to make sure that it got set correctly, and there it is.

[02:26](javascript:;)So now I'm going to download a file, which is just a publicly available Hadoop documentation, HTML file, and if I do an ls, I

[02:35](javascript:;)can see there's my setup.html and I am now going to copy it a couple times to make a setup two and a setup three.

[02:43](javascript:;)If I do an ls, I should see three files.

[02:46](javascript:;)There they are.

[02:46](javascript:;)So the second task is ACLs.

[02:49](javascript:;)We're going to copy this file into the bucket and then configure the access control list for it.

[02:55](javascript:;)So the first one is gsutil command, where I am copying setup.html into my bucket.

[03:06](javascript:;)Once it's copied, I then want to get the default access list that has been assigned to setup.html, which is based on the bucket because that's how we set it.

[03:16](javascript:;)Then right here, I piped it into acl.txt, and now I'm going to cut that, and we can see all of the permissions that had been assigned.

[03:25](javascript:;)So now I want to set the permissions to private.

[03:28](javascript:;)So I'm going to set it to private, and then in order to see it, I'm going to

[03:33](javascript:;)pipe it into acltwo.txt, and then cut that file, and you can see it's now set to private.

[03:39](javascript:;)Update the access list to make the file publicly readable by running the following command, and

[03:45](javascript:;)then I'm going to pipe it into aclthree, so that I can verify what that looks like.

[03:50](javascript:;)You can see it is readable by all users.

[03:53](javascript:;)This is another check point in a lab where you can hit check my progress and in this case is checking if you properly made that file publicly readable.

[04:02](javascript:;)So now I'm going to verify in my bucket using the console that my file is there and that is publicly viewable,

[04:10](javascript:;)and you can tell that based on this little icon and the public link that says that it's accessible to the public.

[04:19](javascript:;)So now in Cloud Shell, I'm going to remove the setup.html in my local Cloud Shell Instance.

[04:27](javascript:;)There it is.

[04:27](javascript:;)Let me remove it from the search here.

[04:31](javascript:;)If I do an ls, I'll see setup two and setup three, but not setup.

[04:35](javascript:;)You can see it got deleted.

[04:38](javascript:;)Let's say I accidentally deleted that from my Cloud Shell Instance, but now I want the copy that was in the bucket back on my local Cloud Shell.

[04:46](javascript:;)So I could just copy from the bucket to my local Cloud Shell, and if I do an ls again, I'll see all three setup files.

[04:54](javascript:;)There they are.

[04:56](javascript:;)The third task is to generate a customer supplied encryption key.

[05:01](javascript:;)To create the key, I'm going to run this command, and that's going to give me some output, and then I can copy this.

[05:11](javascript:;)But first, I'm going to see if I have a boto file.

[05:16](javascript:;)I'm going to do ls-al, and I do not see a boto file.

[05:23](javascript:;)So what I'm going to do is I'm going to run gsutilconfig-n, and then I'm going to do ls-al, and I should now see a boto file.

[05:39](javascript:;)There it is.

[05:40](javascript:;)So I'm going to do a nano.boto, and then I'm going to find the encryption key field, which

[05:49](javascript:;)I'm going to exit back out because I didn't not copy the key that I created, which I need.

[05:57](javascript:;)That is right here.

[05:59](javascript:;)Let me copy that, and let me go back to nano, and let me find the line with encryption underscore key.

[06:20](javascript:;)Could need to expand this because it's very hard to see.

[06:50](javascript:;)See decryption key here is encryption key.

[06:53](javascript:;)I'm going to uncomment this, and then I'm going to paste in my key here.

[07:02](javascript:;)I'm going to press control l, write that file, and then control x to exit nano.

[07:12](javascript:;)So now that I've set that up, I am going to upload the remaining setup two and setup three into the bucket.

[07:20](javascript:;)There's one, and there's the other.

[07:25](javascript:;)Now back in the console, let's go down, I'm going to refresh the bucket.

[07:34](javascript:;)I can see both of these files, and it shows that they are encrypted by a customer supply key.

[07:41](javascript:;)So this is another opportunity to check my progress and make sure I got the points for doing that step.

[07:48](javascript:;)Now what I'm going to do is I am deleting my local files by running remove setup star.

[07:57](javascript:;)So it's going to delete setup, setup two, and setup three.

[08:01](javascript:;)Now I am going to copy the files down from the bucket again, and if I want to cut the encrypted files to see

[08:14](javascript:;)whether I need them back, you can see there they are, and I successfully was able to bring them back even though they're encrypted.

[08:25](javascript:;)So now I'm going to move the current customer supplied encryption key to the decrypt key.

[08:34](javascript:;)So let's go to nano.boto.

[08:39](javascript:;)I'm going to find the comment out the line that I added earlier.

[08:46](javascript:;)I should've noted the line number, so that I wouldn't have to find it again.

[08:54](javascript:;)Decrypt keys in the GSUtil section.

[08:59](javascript:;)Let's see.

[09:11](javascript:;)I think I'm close.

[09:22](javascript:;)I'm looking for that line.

[09:25](javascript:;)So I'm going to comment out encryption key line and uncomment decryption key one right there.

[09:40](javascript:;)Then, I'm going to copy this into decryption key one, and then we save x.

[10:02](javascript:;)So a best practices you would actually delete the old customer key from the encryption line.

[10:10](javascript:;)But in this case, we just copy pasted it.

[10:12](javascript:;)So it's not a big deal.

[10:13](javascript:;)So I'm going generate a new key and then I'm going back to boto files.

[10:24](javascript:;)So I am going to add a new encryption key line, make sure that I copied the new key I made, and then do the same thing again.

[11:08](javascript:;)Sparsed it so I am adding a new encryption key equals, and I'll paste in the new key.

[11:20](javascript:;)Then control O to save, control X to exit.

[11:25](javascript:;)Now, I'm going to rewrite the key for file 1, and comment out the old decrypt key.

[11:41](javascript:;)Again, to bottom.

[11:43](javascript:;)Then I am going to comment out the decryption key 1.

[11:55](javascript:;)Now, while the instructions have you using nano, you definitely could use the Cloud Shell editor as well.

[12:02](javascript:;)That might be a little more pleasant than using this tool, but I'll leave it to you.

[12:09](javascript:;)You would just access that by hitting this little pencil here.

[12:15](javascript:;)It's fine.

[12:15](javascript:;)Decryption key 1 real quick.

[12:19](javascript:;)So we're commenting that out.

[12:22](javascript:;)Then, we're going to save it, and click.

[12:38](javascript:;)Now, we're going to download setup 2, and download setup 3.

[12:55](javascript:;)What happened, no decryption key matches because we commented it out, which makes sense.

[13:03](javascript:;)So the last task in this lab is, we are going to run the following command to view the current life-cycle policy.

[13:09](javascript:;)So we're going to do this.

[13:14](javascript:;)It says it has no life-cycle configuration.

[13:17](javascript:;)So I'm going to create a JSON lifecycle policy file.

[13:23](javascript:;)I'm going to paste the following rule in here.

[13:28](javascript:;)So it's saying if it's over 31 days I'm going to delete it.

[13:34](javascript:;)Writing exit.

[13:37](javascript:;)Then, to set the policy I'm going to run the command provided in the box, and to verify that the policy worked, I'm going to press that.

[13:52](javascript:;)This is another opportunity for you to check your progress and get more points in the lab.

[13:58](javascript:;)This point you should have about 20 out of 35 points.

[14:03](javascript:;)The task 6 is enabling versioning and you can do that by using the following command.

[14:12](javascript:;)Says it suspended, which means it's not enabled.

[14:16](javascript:;)So if we want to enable versioning, we're going to run this command.

[14:22](javascript:;)Then if we were to run the Get-Command again, we would not see that it was suspended we would say that it was enabled.

[14:29](javascript:;)There it is.

[14:31](javascript:;)So check your progress again you'll get more points.

[14:35](javascript:;)The next step, we're going to create several versions of the sample file in the bucket.

[14:40](javascript:;)So I'm going do an ls here.

[14:42](javascript:;)Going to open the setup HTML file.

[14:47](javascript:;)Delete any five lines to change the size.

[14:52](javascript:;)So I'm going to comment out this link and then I'm going to delete all of these links.

[15:14](javascript:;)Probably a faster way to do this thing just holding down delete.

[15:17](javascript:;)That's what I'm doing here.

[15:19](javascript:;)I'm going to delete it all the way to the banner.

[15:27](javascript:;)So I have now effectively changed the size of the file.

[15:32](javascript:;)So I'm going to control O, enter, control X. I'm going to copy the file to the bucket.

[15:42](javascript:;)I'm going to go back to setup.html, delete another five lines.

[15:58](javascript:;)Let's delete some more links.

[16:18](javascript:;)I'm just going to delete up to here.

[16:22](javascript:;)I'm going to save it.

[16:24](javascript:;)Then I'm going to copy it again.

[16:30](javascript:;)So if I wanted to list all versions of the file, which each subsequent one

[16:34](javascript:;)I was deleting different lines and making the size smaller, I was creating a new version.

[16:42](javascript:;)You can see there are three versions: the original one, the one where I deleted

[16:45](javascript:;)the first five lines, and then the one where I deleted the next set of lines.

[16:51](javascript:;)So I am now going to store the version value in the environment variables.

[16:57](javascript:;)So I'm going to say, export version name equals, the oldest version is this one.

[17:10](javascript:;)I'm going to copy that.

[17:11](javascript:;)I'm going to set this variable here make sure it got set correctly, and it is set correctly.

[17:20](javascript:;)Now, I'm going to download the oldest version, call it recovery.text.

[17:30](javascript:;)Then I'm going to verify recovery with a couple of commands.

[17:36](javascript:;)It is saying, c ls setup.html.

[17:54](javascript:;)Looks like that piece didn't work.

[17:59](javascript:;)I think what I did was I set the version name to the wrong thing, it should have been here.

[18:21](javascript:;)So now I can do the Gsutil again, and it still didn't match.

[18:36](javascript:;)So you do this, it's because you didn't follow instructions like me.

[18:41](javascript:;)You should have copied the entire URL for that object.

[18:50](javascript:;)Usually, what happens with the lab is if you have an issue is usually not that the lab is broken.

[18:57](javascript:;)It's usually that you missed a step.

[18:58](javascript:;)So go back three steps and repeat, and that usually works, because you can see here that just worked.

[19:04](javascript:;)Ls.al Setup.html, there's the file.

[19:11](javascript:;)I want to see the recovered text.

[19:15](javascript:;)You can see that the size is different here.

[19:18](javascript:;)So task 7, we're going to synchronize a directory to a bucket and just copy these in.

[19:26](javascript:;)Then I'm going to sync the first-level directory on the VM with my bucket.

[19:38](javascript:;)I'm going to verify that versioning was enabled.

[19:46](javascript:;)How I can check in the browser, I'm going to refresh the bucket, and we go back here, first level.

[19:58](javascript:;)You can see there's a second level.

[20:01](javascript:;)We can see the same thing in the console as we do in the command line.

[20:09](javascript:;)So I can exit Cloud Shell.

[20:12](javascript:;)So now we're going to do some cross-project sharing, this is the last little piece of this lab.

[20:18](javascript:;)So I'm going to open another tab.

[20:21](javascript:;)I'm also going to go to console.cloud.google.com, and I am now signed in, I'm going to select the other project.

[20:38](javascript:;)This one I have 26.

[20:52](javascript:;)I am going to copy the project from the Qwiklabs site, I'm here at lab guide, and I'm going to select that project.

[21:03](javascript:;)This is my other project.

[21:08](javascript:;)Then I am going to now create a bucket for this project.

[21:15](javascript:;)There shouldn't be one in here because it's a new project.

[21:18](javascript:;)I'm going to also call it myproj in project ID, and Create.

[21:32](javascript:;)This will now be bucket name two.

[21:36](javascript:;)So I'm going to upload a file, any file.

[21:53](javascript:;)I've uploaded a screenshot, and this will be my file name.

[22:00](javascript:;)So I'm actually going to rename it.

[22:03](javascript:;)I can't.

[22:10](javascript:;)Now I'm going to go to IAM, service accounts, and then create a service account.

[22:21](javascript:;)I'm going to call it cross-project-storage, and click create.

[22:29](javascript:;)Then I'm going to give it the Storage Object Viewer.

[22:41](javascript:;)Click continue.

[22:44](javascript:;)I'm going to create a key.

[22:46](javascript:;)I'm going to select JSON, create, and then it's going to download that file for me.

[22:53](javascript:;)It's there.

[22:55](javascript:;)Hit close, and I can hit done.

[23:01](javascript:;)So I am now going to rename this credentials.json.

[23:19](javascript:;)Here it is. I'm going to switch back to the other project, check my progress, and I should get five more points.

[23:32](javascript:;)So now we are just five points away from finishing the lab.

[23:35](javascript:;)Now we are in project ID one, and we're going to create a VM, create.

[23:49](javascript:;)Calling it crossproject.

[23:53](javascript:;)I'm going to make it in Europe in D, and I'm making it a micro, and create.

[24:11](javascript:;)VM is ready, I'm going to SSH into it.

[24:17](javascript:;)There it is, click SSH, then move my window back here to get the bucket name of the project I created here.

[25:00](javascript:;)I'm going to verify that it worked, and then I'm going to export the file name of the file that I uploaded.

[25:17](javascript:;)Grab that, put quotes around there because that space is in it.

[25:43](javascript:;)Verify that worked, and there it is.

[25:52](javascript:;)LS what's in that bucket.

[25:55](javascript:;)Form a VM on this side, it tells me that I don't have access to do that, so now I'm going to verify that.

[26:03](javascript:;)I am going to upload here, upload file.

[26:08](javascript:;)I'm going to select the credentials.json that I downloaded.

[26:16](javascript:;)Close, then I'm going to authorize that file to verify access.

[26:28](javascript:;)I'm going to do this again, and now I can see my file in there.

[26:35](javascript:;)I can do it with the file as well.

[26:40](javascript:;)Let me try to copy these credentials so they don't have access to that project.

[26:56](javascript:;)So if I wanted to do that, I would go back to this project, and modify the role in IAM.

[27:02](javascript:;)It Should be my last step here, going back to IAM, cross-project-storage, pencil.

[27:18](javascript:;)I'm also going to give Storage Object Admin, save.

[27:30](javascript:;)Once I hit save, I can check my progress, and then you will have all of the points in the lab.

[27:41](javascript:;)The last step is optional, you're just going to return to your SSH terminal, and

[27:46](javascript:;)verify that everything is good to go but that is the entire walkthrough for this lab.

[27:52](javascript:;)I hope you enjoyed it.

Cloud SQL-

[00:00](javascript:;)Let’s dive into the structured or relational database services.

[00:03](javascript:;)First up is Cloud SQL.

[00:05](javascript:;)Why would you use a Google Cloud service for SQL, when you can install a SQL Server application image on a VM using Compute Engine?

[00:13](javascript:;)The question really is, should you build your own database solution or use a managed service?

[00:20](javascript:;)There are benefits to using a managed service, so let’s learn about why you’d use Cloud SQL as a managed service inside of Google Cloud.

[00:27](javascript:;)Cloud SQL is a fully managed service of either MySQL, PostgreSQL, or Microsoft SQL Server databases.

[00:38](javascript:;)This means that patches and updates are automatically applied, … but you still have to administer MySQL users with the native authentication tools that come with these databases.

[00:48](javascript:;)Cloud SQL supports many clients, such as Cloud Shell, App Engine and Google Workspace scripts.

[00:55](javascript:;)It also supports other applications and tools that you might be used to like SQL Workbench, Toad and other external applications using standard MySQL drivers.

[01:06](javascript:;)Cloud SQL delivers high performance and scalability with up to 64 TB of storage capacity, 60,000 IOPS, and 624 GB of RAM per instance.

[01:18](javascript:;)You can easily scale up to 96 processor cores and scale out with read replicas.

[01:23](javascript:;)Currently, you can use Cloud SQL with either MySQL 5.6, 5.7, or 8.0, PostgreSQL 9.6, 10, 11, 12,

[01:32](javascript:;)13, or 14, or either of the Web, Express, Standard or Enterprise SQL Server 2017 or 2019 editions.

[01:42](javascript:;)Let’s focus on some other services provided by Cloud SQL: In HA configuration, within a regional instance, the configuration is made up of a primary instance and a standby instance.

[01:55](javascript:;)Through synchronous replication to each zone's persistent disk, all writes made to the primary instance are replicated to disks in both zones before a transaction is reported as committed.

[02:06](javascript:;)In the event of an instance or zone failure, the persistent disk is attached to the standby instance, and it becomes the new primary instance.

[02:14](javascript:;)Users are then rerouted to the new primary.

[02:16](javascript:;)This process is called a failover.

[02:20](javascript:;)Cloud SQL also provides automated and on-demand backups with point-in-time recovery.

[02:26](javascript:;)You can import and export databases using mysqldump, or import and export CSV files.

[02:31](javascript:;)Cloud SQL can also scale up, which does require a machine restart or scale out using read replicas.

[02:39](javascript:;)That being said, if you are concerned about horizontal scalability, you’ll want to consider Cloud Spanner which we’ll cover later in this module.

[02:48](javascript:;)Choosing a connection type to your Cloud SQL instance will affect how secure, performant, and automated it will be.

[02:56](javascript:;)If you’re connecting an application that is hosted within the same Google Cloud project as your Cloud SQL instance, and it is

[03:03](javascript:;)collocated in the same region, choosing the Private IP connection will provide you with the most performant and secure connection using private connectivity.

[03:14](javascript:;)In other words, traffic is never exposed to the public internet.

[03:18](javascript:;)Note that connecting to the Cloud SQL Private IP address from VMs in the same region is only a performance-based recommendation and not a requirement.

[03:29](javascript:;)If the application is hosted in another region or project, or if you are trying

[03:33](javascript:;)to connect to your Cloud SQL instance from outside of Google Cloud, you have 3 options.

[03:40](javascript:;)In this case, I recommend using the Cloud SQL Auth Proxy, which handles authentication, encryption, and key rotation for you.

[03:50](javascript:;)If you need manual control over the SSL connection, you can generate and periodically rotate the certificates yourself.

[03:57](javascript:;)Otherwise, you can use an unencrypted connection by authorizing a specific IP address to connect to your SQL server over its external IP address.

[04:08](javascript:;)You will explore these options in an upcoming lab.

[04:12](javascript:;)To summarize, let’s explore this decision tree to help you find the right data storage service with full relational capability.

[04:20](javascript:;)Memorystore provides a fully-managed in-memory data store service for workloads requiring microsecond response times, or that have large spikes in traffic, as seen in gaming environments and real-time analytics.

[04:33](javascript:;)If you don’t need an in-memory data store, but your use case is relational data used primarily for analytics, these workloads are best supported by BigQuery.

[04:42](javascript:;)However, if your relational data workload isn’t analytics the choice lies between Cloud Spanner and Cloud SQL.

[04:50](javascript:;)If you don’t need horizontal scaling or a globally available system, Cloud SQL is a cost-effective solution.

Lab Intro: Cloud SQL-

[00:00](javascript:;)Let's take some of the Cloud SQL concepts that we just discussed, and apply them in a lab.

[00:05](javascript:;)In this lab you configure a Cloud SQL Server, and learn how to connect an application to it via a proxy over an external connection.

[00:12](javascript:;)You also configure a connection over a private IP link that offers performance and security benefits.

[00:19](javascript:;)The app we chose to demonstrate in this lab is WordPress, but the information and best practices are applicable to any application that needs an SQL Server.

[00:29](javascript:;)By the end of this lab, you will have two working instances of a WordPress front end

[00:33](javascript:;)connected over two different connection types to its SQL instance back end, as shown in this diagram.

Implementing Cloud SQL- Lab to be done

Lab Review: Cloud SQL-

[00:00](javascript:;)In this lab you created a Cloud SQL database and configured it to use both an

[00:04](javascript:;)external connection over a secure proxy and a private IP address, which is more secure and performant.

[00:10](javascript:;)If your application is hosted in another Region, VPC, or even project, use a proxy to secure its connection over the external connection.

[00:20](javascript:;)You can stay for a lab walkthrough, but remember that GCP's user interface can change, so your environment might look slightly different.

[00:28](javascript:;)Welcome to the lab walkthrough for implementing Cloud SQL.

[00:32](javascript:;)At this point in the lab, I have logged in with the username and password that QwikLabs has provided me.

[00:39](javascript:;)My first task is to create a Cloud SQL database.

[00:42](javascript:;)And go here to SQL, And hit Create Instance.

[00:51](javascript:;)I'm going to choose my SQL, And I'm going to call this wordpress-db.

[01:04](javascript:;)For password I'm just going to use the word password, so that I make sure I'm not forgetting it.

[01:09](javascript:;)I recommend you use something very simple.

[01:11](javascript:;)I'm going to use US Central 1.

[01:14](javascript:;)I am going to expand the configuration options, and then in Connectivity I'm going to select the private IP.

[01:22](javascript:;)Hit Enable API, and once that's been enabled, which could take a couple of seconds, I'm going to hit Allocate and Connect.

[01:31](javascript:;)And just so you know, this could take three to five minutes, so just be patient.

[01:37](javascript:;)Once that's done, it is going to make this Create button enabled.

[01:42](javascript:;)Feel free to look through some of the other things that it's calling out in the lab, like configuring the machine type and changing the storage capacity.

[01:53](javascript:;)If you add a couple of zeros, you can see the throughput increases.

[01:59](javascript:;)Set it back to 10, hit Close here.

[02:03](javascript:;)Again, this could take three to five minutes, so be patient.

[02:06](javascript:;)Once it's done, it'll say Create here.

[02:13](javascript:;)All right, so my IP has been allocated, and now I can hit Create.

[02:20](javascript:;)Here we go.

[02:21](javascript:;)And now, that took a while, but creating your Cloud SQL instance or Cloud SQL instance, might take even longer.

[02:30](javascript:;)So be patient, but while this is creating I can do other steps in the lab.

[02:36](javascript:;)The verification in step 15 is going to require that your Cloud SQL instance is running, and that there is a green check here.

[02:45](javascript:;)So you won't be able to get that step and those five points until this is done, but you can do some of the other steps while we wait.

[02:53](javascript:;)So while this is going, I'm going to open another tab so that this is still running and I can check on it.

[03:00](javascript:;)And I am now going to go to Compute Engine, which I could either go here, or this tile right here has my Compute Engine instances.

[03:08](javascript:;)As you can see, two have been created for me. wordpress-europe-proxy, which is the proxy for my Cloud SQL instance and for the private IP instance.

[03:21](javascript:;)So for this one, I'm going to click SSH.

[03:27](javascript:;)And when that is ready, to SSH into I am going to download the Cloud SQL proxy and then I'm going to make it executable.

[03:35](javascript:;)So I'm going to do a wget, Enter.

[03:38](javascript:;)It's copied it and made it executable.

[03:42](javascript:;)So in order to start the proxy, you need the connection name of the Cloud SQL instance, which requires that it's actually running.

[03:50](javascript:;)So we'll go ahead, and go back here, and see what the status is. I'm going to refresh and see if anything is there.

[04:01](javascript:;)Now it's not there, but it does let me click on it.

[04:03](javascript:;)And let's see if the connection name is there, and it is here.

[04:06](javascript:;)Instance connection name, so I'm going to copy that, and I'm going to go back to my SSH window.

[04:15](javascript:;)And I am going to create an environment variable for that connection instance.

[04:22](javascript:;)So I'm going to do an export SQL\_CONNECTION =, paste that in there.

[04:34](javascript:;)And if I want to verify that that environment variable is set, I'm going to do an echo SQL, oops, spell it right.

[04:45](javascript:;)And it should output that, and there it is.

[04:49](javascript:;)So I'm not going to get too far ahead, because the rest of the steps do require that my Cloud SQL instance is running.

[04:58](javascript:;)You can see it's still creating, so it is not going to let me create a database yet.

[05:04](javascript:;)So I'm just going to leave it on this screen, so that when it is completely done it'll let me create a database, which I need for the next step.

[05:13](javascript:;)All right, at this point my instance is running.

[05:16](javascript:;)So if I go to Cloud SQL, it'll have a green checkmark next to wordpress-db.

[05:22](javascript:;)Right here, it says it's runnable.

[05:23](javascript:;)So the step I was missing is I need to create a database, and I am going

[05:27](javascript:;)to create a database called wordpress, because that's what the application expects, and I'm going to hit Create.

[05:35](javascript:;)And now I am going to return to, My SSH window, and I'm just going to make sure that it still has my environment variable.

[05:48](javascript:;)Then I am going to activate the proxy connection to my SQL database.

[05:53](javascript:;)By running this, I'll run it in the background.

[05:57](javascript:;)And then I'm going to expect that it says Ready for new connections, which it has output.

[06:03](javascript:;)I'm going to press Enter.

[06:05](javascript:;)And this is a point in the lab where you can also hit Check my Progress.

[06:09](javascript:;)And at this point you should have all ten points in the lab, but we're still going to do one more step, which is task 3.

[06:17](javascript:;)Actually, task 3 and task 4, where we're going to connect the application to the Cloud SQL instance.

[06:23](javascript:;)So I'm going to configure the wordpress application.

[06:27](javascript:;)So I'm going to copy the curl command.

[06:30](javascript:;)Go ahead and run that, and it is going to output the external IP address for my virtual machine.

[06:37](javascript:;)I'm going to copy that, and open it here, And then I'm going to hit Let's Go.

[06:46](javascript:;)I am going to leave everything with default, except I'm going to change the username

[06:50](javascript:;)to root and I'm going to put in the password that I defined, which is password.

[06:55](javascript:;)And then for database host I am going to use a localhost IP, which is 127.0.0.1, and then I'm going to hit Submit.

[07:09](javascript:;)Now, when the connection has been made, it is going to let me install WordPress, and I'm going to click Run the Installation.

[07:16](javascript:;)This could take a few moments to complete.

[07:18](javascript:;)Once this is done I should get a success window, and this can take up to three minutes depending on where you are running your lab from.

[07:28](javascript:;)So here it goes, apparently it's very fast for me.

[07:32](javascript:;)And so once the connection has been made, I am going to go here, and I am going to remove everything passed to the external IP.

[07:42](javascript:;)Delete, hit Enter.

[07:48](javascript:;)And when this loads, I should be able to see my blog.

[08:00](javascript:;)So it looks like it's still installing, so I will go ahead and be patient.

[08:04](javascript:;)Add some Information that I don't need to remember.

[08:09](javascript:;)So this is My Fake Site title.

[08:13](javascript:;)Any username, leave that password it gave me, m@b.com, and Install.

[08:23](javascript:;)And this was actually a step in the lab that I ignored, which was step 7, so don't do what I did and skip a step.

[08:37](javascript:;)So I had a successful installation.

[08:39](javascript:;)So what I'm going to do here is I'm going to remove all of the information that's after the external IP.

[08:44](javascript:;)When I hit Enter, it should take me to my blog.

[08:47](javascript:;)And Hello world!

[08:50](javascript:;)This is my blog, success.

[08:52](javascript:;)So the last task is to connect to Cloud SQL via the internal IP.

[08:58](javascript:;)So I'm going to go back here, and go to SQL and gcp.

[09:04](javascript:;)I'm going to click on wordpress-db, and then I am going to note the private IP address here.

[09:11](javascript:;)And I'm actually going to note it on a note, because pretty sure that it's going to have me copy something else.

[09:20](javascript:;)So make sure you copy it down somewhere in a clipboard, and then I'm going to go to Compute Engine.

[09:25](javascript:;)And it is going to want me to copy the external IP address for WordPress private IP.

[09:39](javascript:;)I'm going to copy that, and paste it in a new tab, press Enter.

[09:44](javascript:;)I'm going to hit Let's Go, and then the database name I'm going to leave alone,

[09:49](javascript:;)and I'm going to change this to root, and leave password because that's what I put before.

[09:55](javascript:;)Then I'm going to put the SQL private IP that I copied earlier to my Notepad, just need to find it.

[10:06](javascript:;)It's here, I'm going to copy that in here and hit Submit.

[10:12](javascript:;)And then I'm going to hit Run the Installation, and I should get Already Installed.

[10:19](javascript:;)So I created a direct connection to a private IP instead of configuring a proxy, and that connection is private.

[10:28](javascript:;)If I remove here, the same private IP should get my blog, and there it is.

[10:37](javascript:;)So in review, we created a Cloud SQL database and we configured it to use an external connection over a secure proxy as well as a private IP address.

[10:46](javascript:;)Hope you enjoyed the lab, thanks for watching.

Cloud Spanner-

[00:00](javascript:;)If cloud SQL does not fit your requirements because you need horizontal scalability, consider using cloud spanner.

[00:07](javascript:;)Cloud spanner is a service built for the cloud specifically to combine the benefits of relational database structure with non relational horizontal scale.

[00:16](javascript:;)This service can provide petabytes of capacity and offers transactional consistency at global scale schemas, SQL and automatic synchronous replication for high availability.

[00:29](javascript:;)Use cases include financial applications and inventory applications traditionally served by relational database technology depending on whether you

[00:37](javascript:;)create a multi regional or regional instance, you'll have different monthly up time sls as shown on this slide.

[00:45](javascript:;)However, for up to date numbers, you should always refer to the documentation, which you'll find in the link section of this video.

[00:52](javascript:;)Let's compare cloud spanner with both relational and non relational databases like a relational database.

[00:58](javascript:;)Cloud spanner has schema, SQL and strong consistency.

[01:03](javascript:;)Also like a non relational database, Cloud spanner offers high availability, horizontal scalability and configurable replication, as mentioned, Cloud spanner offers the best of the relational and non relational worlds.

[01:19](javascript:;)These features allow for mission critical use cases such as building consistent systems for transactions and inventory management in the financial services and retail industries.

[01:30](javascript:;)To better understand how all of it works.

[01:32](javascript:;)Let's look at the architecture of cloud spanner.

[01:35](javascript:;)A cloud spanner instance replicates data in end cloud zones which can be within one region or across several regions.

[01:43](javascript:;)The database placement is configurable, meaning you can choose which region to put your database in.

[01:49](javascript:;)This architecture allows for high availability and global placement.

[01:53](javascript:;)The replication of data will be synchronized across zones using Google's global fiber network.

[01:58](javascript:;)Using atomic clocks ensures adamiscity whenever you are updating your data.

[02:03](javascript:;)That's as far as we're going to go with cloud spanner.

[02:05](javascript:;)Because the focus of this module is to understand the circumstances when you would use cloud spanner, let's look at a decision tree.

[02:13](javascript:;)If you have outgrown any relational database are shutting your databases for throughput, high performance, need transactional consistency, global data and strong consistency.

[02:24](javascript:;)Or just want to consolidate your database.

[02:26](javascript:;)Consider using cloud spanner.

[02:28](javascript:;)If you don't need any of these nor full relational capabilities, consider a no SQL service, such as cloud Fire store, which we will cover next.

[02:37](javascript:;)If you're now convinced that using cloud spanner as a managed service is better than using or re implementing your

[02:44](javascript:;)existing my SQL solution, see the link section for a solution on how to migrate from my SQL to cloud spanner.

Firestore-

[00:00](javascript:;)If you are looking for a highly scalable NoSQL database for your applications, consider using Cloud Firestore.

[00:07](javascript:;)Cloud Firestore is a fast, fully managed, serverless, cloud native, NoSQL, document database that simplifies storing, synking and querying data for your mobile web and IOT apps at global scale.

[00:22](javascript:;)Its client libraries provide live synchronization and offline support and it's security features and integrations with Firebase and GCP accelerate building truly serverless apps.

[00:33](javascript:;)Cloud Firestore also supports ACID transactions so if any of the operations in the transaction fail and cannot be retried, the whole transaction will fail.

[00:44](javascript:;)Also with automatic multi region replication and strong consistency, your data is safe and available even when disasters strike.

[00:53](javascript:;)Cloud Firestore even allows you to run sophisticated queries against your NoSQL data without any degradation in performance.

[01:01](javascript:;)This gives you more flexibility in the way you structure your data.

[01:05](javascript:;)Cloud Firestore is actually the next generation of Cloud Datastore.

[01:08](javascript:;)Cloud Firestore can operate in Datastore mode, making it backwards compatible with Cloud Datastore.

[01:15](javascript:;)By creating a Cloud Firestore database in Datastore mode, you can access Cloud Firestore's improveD storage layer while keeping Cloud Datastore system behavior.

[01:23](javascript:;)This removes the following Cloud Datastore limitations.

[01:27](javascript:;)Queries are no longer eventually consistent instead, they are all strongly consistent.

[01:33](javascript:;)Transactions are no longer limited to 25 entity groups, rights to an entity group are no longer limited to 1 per second.

[01:42](javascript:;)Cloud Firestore in native mode introduces new features such as a new, strongly consistent storage layer, a collection and document data model, real time updates, mobile and web client libraries.

[01:57](javascript:;)Cloud Firestore is backward compatible with Cloud Datastore but the new data model, real time updates in mobile and web client library features are not.

[02:06](javascript:;)To access all of the new Cloud Firestore features, you must use Cloud Firestore in native mode.

[02:12](javascript:;)A general guideline is to use Cloud Firestore in Datastore mode for new server projects and native mode for new mobile and web apps.

[02:21](javascript:;)As the next generation of Cloud Datastore, Cloud Firestore is compatible with all Cloud Datastore, APIs and client libraries.

[02:29](javascript:;)Existing Cloud Datastore users will be live upgraded to Cloud Firestore automatically at a future date.

[02:36](javascript:;)For more information, see the link section of this video, to summarize, let's explore this decision tree.

[02:42](javascript:;)To help you determine whether Cloud Firestore is the right storage service for your data.

[02:47](javascript:;)If your schema might change and you need an adaptable database, you need to scale

[02:51](javascript:;)to zero or you want low maintenance overhead scaling up to terabytes consider using Cloud Firestore.

[02:58](javascript:;)Also, if you don't require transactional consistency, you might want to consider Cloud Bigtable.

[03:04](javascript:;)Depending on the cost or size, I will cover Cloud Bigtable next.

Cloud Bigtable-

[00:00](javascript:;)If you don’t require transactional consistency, you might want to consider Cloud Bigtable.

[00:05](javascript:;)Cloud Bigtable is a fully managed NoSQL database with petabyte-scale and very low latency.

[00:12](javascript:;)It seamlessly scales for throughput and it learns to adjust to specific access patterns.

[00:17](javascript:;)Cloud Bigtable is actually the same database that powers many of Google’s core services, including Search, Analytics, Maps, and Gmail.

[00:27](javascript:;)Cloud Bigtable is a great choice for both operational and analytical applications, including IoT, user

[00:32](javascript:;)analytics, and financial data analysis, because it supports high read and write throughput at low latency.

[00:39](javascript:;)It’s also a great storage engine for machine learning applications.

[00:44](javascript:;)Cloud Bigtable integrates easily with popular big data tools like Hadoop, Cloud Dataflow, and Cloud Dataproc.

[00:51](javascript:;)Plus, Cloud Bigtable supports the open source industry standard HBase API, which makes it easy for your development teams to get started.

[01:00](javascript:;)Cloud Dataflow and Cloud Dataproc are covered late in the course series.

[01:04](javascript:;)For more information on the HBase API, see the links section of this video.

[01:10](javascript:;)Cloud Bigtable stores data in massively scalable tables, each of which is a sorted key/value map.

[01:16](javascript:;)The table is composed of rows, each of which typically describes a single entity, and columns, which contain individual values for each row.

[01:26](javascript:;)Each row is indexed by a single row key, and columns that are related to one another are typically grouped together into a column family.

[01:35](javascript:;)Each column is identified by a combination of the column family and a column qualifier, which is a unique name within the column family.

[01:44](javascript:;)Each row/column intersection can contain multiple cells, or versions, at different timestamps, providing a record of how the stored data has been altered over time.

[01:56](javascript:;)Cloud Bigtable tables are sparse; if a cell does not contain any data, it does not take up any space.

[02:03](javascript:;)The example shown here is for a hypothetical social network for United States presidents, where each president can follow posts from other presidents.

[02:11](javascript:;)Let me highlight some things: \* The table contains one column family, the follows family.

[02:17](javascript:;)This family contains multiple column qualifiers.

[02:21](javascript:;)\* Column qualifiers are used as data.

[02:23](javascript:;)This design choice takes advantage of the sparseness of Cloud Bigtable tables, and the fact that new column qualifiers can be added as your data changes.

[02:32](javascript:;). \* The username is used as the row key.

[02:36](javascript:;)Assuming usernames are evenly spread across the alphabet, data access will be reasonably uniform across the entire table.

[02:45](javascript:;)This diagram shows a simplified version of Cloud Bigtable’s overall architecture.

[02:50](javascript:;)It illustrates that processing, which is done through a front-end server pool and nodes, is handled separately from the storage.

[02:57](javascript:;)A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries.

[03:05](javascript:;)Tablets are similar to HBase regions, for those of you who have used the HBase API.

[03:12](javascript:;)Tablets are stored on Colossus, which is Google's file system, in SSTable format.

[03:17](javascript:;)An SSTable provides a persistent, ordered immutable map from keys to values, where both keys and values are arbitrary byte strings.

[03:27](javascript:;)As I mentioned earlier, Cloud Bigtable learns to adjust to specific access patterns.

[03:32](javascript:;)If a certain Bigtable node is frequently accessing a certain subset of data... … Cloud Bigtable

[03:37](javascript:;)will update the indexes so that other nodes can distribute that workload evenly, as shown here.

[03:43](javascript:;)That throughput scales linearly, so for every single node that you do add, you're going to see a linear scale of throughput performance, up to hundreds of nodes.

[03:53](javascript:;)In summary, if you need to store more than 1 TB of structured data, have very high volume of writes, need read/write latency of

[04:01](javascript:;)less than 10 milliseconds along with strong consistency, or need a storage service that is compatible with the HBase API, consider using Cloud Bigtable.

[04:12](javascript:;)If you don’t need any of these and are looking for a storage service that scales down well, consider using Firestore.

[04:19](javascript:;)Speaking of scaling, the smallest Cloud Bigtable cluster you can create has three nodes and can handle 30,000 operations per second.

[04:28](javascript:;)Remember that you pay for those nodes while they are operational, whether your application is using them or not.

Memorystore-

[00:00](javascript:;)Let's get a quick overview of Memorystore.

[00:01](javascript:;)Memorystore for Redis provides a fully managed in-memory data store service built on scalable, secure and highly available infrastructure managed by Google.

[00:13](javascript:;)Applications running on Google Cloud can achieve extreme performance by leveraging the highly scalable, available, secure Redis service without the burden of managing complex Redis deployments.

[00:25](javascript:;)This allows you to spend more time writing code so that you can focus on building great apps.

[00:32](javascript:;)Memorystore also automates complex tasks like enabling high availability, failover, patching, and monitoring.

[00:40](javascript:;)High availability instances are replicated across two zones and provide a 99.9% availability SLA.

[00:48](javascript:;)You can easily achieve the sub-millisecond latency and throughput your applications need.

[00:52](javascript:;)Start with the lowest tier and smallest size, and grow your instance effortlessly with minimal impact to application availability.

[01:01](javascript:;)Memorystore can support instances up to 300 GB and network throughput of 12 gigabits per second.

[01:09](javascript:;)Because Memorystore for Redis is fully compatible with the Redis Protocol, you can lift and shift

[01:14](javascript:;)your applications from open source Redis to Memorystore without any code changes by using the import/export feature.

[01:22](javascript:;)There's no need to learn new tools because all existing tools in client libraries just work.

Quiz: Storage and Database Services-

Which data storage service provides data warehouse services for storing data but also offers an interactive SQL interface for querying the data?

BigQuery

BigQuery is a data warehousing service that allows the storage of huge data sets while making them immediately processable without having to extract or run the processing in a separate service.

Which Google Cloud data storage service offers ACID transactions and can scale globally?

Cloud Spanner

Cloud Spanner provides ACID (Atomicity, Consistency, Isolation, Durability) properties that enable transactional reads and writes on the database. It can also scale globally.

What data storage service might you select if you just needed to migrate a standard relational database running on a single machine in a datacenter to the cloud?

Cloud SQL

Cloud SQL offers a PostgreSQL server or a MySQL server as a managed service.

Module Review-

[00:00](javascript:;)In this module, we covered the different storage and database services that GCP offers.

[00:05](javascript:;)Specifically, you learned about Cloud Storage, a fully managed object store; Cloud SQL, a fully managed

[00:12](javascript:;)MySQL and PostgreSQL database service; Cloud Spanner, a relational database service with transactional consistency, global scale

[00:22](javascript:;)and high availability; Cloud Fire Store, a fully managed NoSQL document database; Cloud Bigtable, a fully

[00:32](javascript:;)managed NoSQL wide column database and Cloud Memorystore, a fully managed in-memory data store service for Redis.

[00:42](javascript:;)From an infrastructure perspective, the goal was to understand what services are available and how they are used in different circumstances.

[00:50](javascript:;)Defining a complete data strategy is beyond the scope of this course.

[00:54](javascript:;)However, Google offers courses on Data Engineering and Machine Learning on GCP that cover data strategy.

RESOURCE MANAGEMENT:

Module Overview-

[00:01](javascript:;)In this module we will cover resource management.

[00:03](javascript:;)Resources in GCP are billable so managing them means controlling cost.

[00:08](javascript:;)There are several methods in place for controlling access to the resources and there are quotas that limit consumption.

[00:15](javascript:;)In most cases, the default quotas can be raised on request.

[00:18](javascript:;)But having them in place provides a checkpoint or a chance to make sure that this really is a resource you intend to consume in greater quantity.

[00:27](javascript:;)In this module, we will build on what we learned in the cloud IAM module.

[00:32](javascript:;)First, I will provide an overview of the resource manager.

[00:36](javascript:;)Then we will go into quotas, labels, and names.

[00:39](javascript:;)Next we will cover billing to help you set budgets and alerts.

[00:43](javascript:;)To complete your learning experience, you will get to examine billing data with BigQuery in a lab.

[00:50](javascript:;)Let's get started with an overview of resource manager.

Resource Manager-

[00:00](javascript:;)The resource manager lets you hierarchically manage resources by project, folder, and organization.

[00:06](javascript:;)This should sound familiar because we covered it in the Cloud IAM module.

[00:10](javascript:;)Let me refresh your memory.

[00:12](javascript:;)Policies contain a set of roles and members, and policies are set on resources.

[00:17](javascript:;)These resources inherit policies from their parent as we can see on the left.

[00:22](javascript:;)Therefore, resource policies are a union of parent and resource.

[00:27](javascript:;)Also, keep in mind that if a parent policy is less restrictive, it overrides the more restrictive resource policy.

[00:34](javascript:;)Although IAM policies are inherited top to bottom, billing is accumulated from the bottom up, as we can see on the right.

[00:43](javascript:;)Resource consumption is measured in quantities like rate of use or time, number of items, or feature use.

[00:50](javascript:;)Because a resource belongs to only one project, a project accumulates the consumption of all its resources.

[00:56](javascript:;)Each project is associated with one billing account, which means that an organization contains all billing accounts.

[01:03](javascript:;)Let's explore organizations, projects, and resources more.

[01:07](javascript:;)Just to reiterate, an organization node is the root node for all Google Cloud Platform resources.

[01:13](javascript:;)This diagram shows an example where we have an individual Bob, who has control of the organizational domain through the organization admin role.

[01:22](javascript:;)Bob has delegated privileges and access to the individual projects to Alice by making her a project creator.

[01:30](javascript:;)Because a project accumulates the consumption of all its resources, it can be used to track resources and quota usage.

[01:37](javascript:;)Specifically, projects that you enable billing, manage permissions and credentials, and enabled service and APIs.

[01:45](javascript:;)To interact with Cloud Platform resources, you must provide the identifying project information for every request.

[01:51](javascript:;)A project can be identified by the project name, which is a human-readable way to identify your projects, but it isn't used by any Google APIs.

[02:00](javascript:;)There's also the project number, which is automatically generated by the server and assigned to your project,

[02:06](javascript:;)and there is the Project ID, which is a unique ID that is generated from your project name.

[02:12](javascript:;)You can find these three identifying attributes on the dashboard of your GCP console, or by querying the Resource Manager API.

[02:20](javascript:;)Finally, let's talk about the resource hierarchy.

[02:23](javascript:;)From a physical organization standpoint, resources are categorized as global, regional, or zonal.

[02:30](javascript:;)Let's look at some examples.

[02:32](javascript:;)Images, snapshots, and networks, are global resources.

[02:36](javascript:;)External IP addresses are regional resources, and instances and disks are zonal resources.

[02:43](javascript:;)However, regardless of the type, each resource is organized into a project.

[02:48](javascript:;)This enables each project to have its own billing and reporting.

Quotas-

[00:00](javascript:;)Now that we know that a project accumulates the consumption of all its resources, let's talk about quotas.

[00:06](javascript:;)All resources in Google Cloud are subject to project quotas or limits.

[00:12](javascript:;)These typically fall into one of the three categories shown here.

[00:17](javascript:;)How many resources you can a project create?

[00:19](javascript:;)For example, you can only have fifteen VPC networks per project.

[00:23](javascript:;)How quickly you can make API requests in a project or rate limits.

[00:30](javascript:;)For example, by default, you can only make five administrative actions per second per project when using the Cloud Spanner API.

[00:38](javascript:;)And three - Regional quotas For example, by default, you can only have 24 CPUs per region.

[00:48](javascript:;)Given these quotas, you may be wondering, how do I spin up one of those 96 Core VMs.

[00:54](javascript:;)As your use of Google Cloud expands over time, your quotas may increase accordingly.

[00:59](javascript:;)If you expect a notable upcoming increase in usage, you can proactively request quota adjustments from the quotas page in the Cloud console.

[01:08](javascript:;)This page will also display your current quotas.

[01:11](javascript:;)If quotas can be changed, why do they exist?

[01:15](javascript:;)Project quotas prevent runaway consumption in case of error or malicious attack.

[01:20](javascript:;)For example, imagine you accidentally create 100 instead of 10 Compute Engine instances using the G Cloud command line.

[01:28](javascript:;)Quotas also prevent billing spikes or surprises.

[01:32](javascript:;)Quotas are related to billing, but we will go through how to set up budgets and alerts later, which will really help you manage billing.

[01:39](javascript:;)Finally, quotas for sizing consideration and periodic review.

[01:44](javascript:;)For example, do you really need a 96 Core instance?

[01:48](javascript:;)Or can you go with a smaller and cheaper alternative?

[01:52](javascript:;)It is also important to mention the quotas are the maximum amount of resources you can create for that resource type as long as those resources are available.

[02:01](javascript:;)Quotas do not guarantee that resources will be available at all times.

[02:06](javascript:;)For example, if a region is out of local SSDs, you cannot create local SSDs in that region, even if you still hae quota for local SSDs.

Labels-

[00:00](javascript:;)Projects and folders provide levels of segregation for resources.

[00:04](javascript:;)But what if you want more granularity, that's where labels come in.

[00:08](javascript:;)Labels are utility for organizing Google Cloud resources.

[00:12](javascript:;)Labels are key value pairs that you can attach to your resources, like VMs, disks, snapshots and images.

[00:19](javascript:;)You can create and manage labels using the Google Cloud Console, gcloud or the Resource Manager API.

[00:27](javascript:;)Each resource can have up to 64 labels.

[00:31](javascript:;)For example, you could create a label to define the environment of your virtual machines.

[00:36](javascript:;)Then, you define the label for each of your instances as either production or test.

[00:42](javascript:;)Using this label, you could search and list all of your production resources for inventory purposes.

[00:49](javascript:;)Labels can also be used in scripts to help analyze costs or for run bulk operations on multiple resources.

[00:56](javascript:;)The screenshot on the right shows an example of four labels that are created on an instance.

[01:02](javascript:;)Let's go over some examples of what labels are used for.

[01:06](javascript:;)I recommend adding labels based on team or cost center to distinguish instances owned by different teams.

[01:14](javascript:;)You can use this type of label for cost accounting or budgeting, for example, team marketing and team research.

[01:22](javascript:;)You can also use labels to distinguish components, for example, component redis, component frontend.

[01:29](javascript:;)Again, you can label based on environment or stage.

[01:35](javascript:;)You should also consider using labels to define the owner or primary contact for a resource.

[01:41](javascript:;)For example, owner gaurav, contact OPM, or add labels to resources to define their state, for example, state in use, state ready for deletion.

[01:54](javascript:;)Now, it's important not to confuse labels with tags.

[01:58](javascript:;)Labels we just learned are user-defined strings and key-value formats that are used to organize resources and they can propagate true billing.

[02:08](javascript:;)Tags, on the other hand, are user-defined strings that are applied to instances only and are mainly used for networking, such as applying firewall rules.

[02:18](javascript:;)For more information about using labels, see the link section of this video.

Billing-

[00:00](javascript:;)Because the consumption of all resources under a project accumulates into one billing account, let's talk billing.

[00:06](javascript:;)To help with project planning and controlling costs, you can set a budget.

[00:10](javascript:;)Setting a budget lets you track how your spend is growing towards that amount.

[00:15](javascript:;)This screenshot shows the budget creation interface.

[00:18](javascript:;)First, you set a budget name and specify which project this budget applies to.

[00:23](javascript:;)Then you can set the budget at a specific amount or match it to the previous month's spend.

[00:29](javascript:;)After you determine your budget amount, you can set the budget alerts.

[00:33](javascript:;)These alerts send emails to billing admins after spend exceeds a percentage of the budget or a specified amount.

[00:39](javascript:;)In our case, it would send an email when spending reaches 50 percent, 90 percent, and 100 percent of the budget amount.

[00:46](javascript:;)You can even choose to send an alert when the spend is forecasted to exceed the percent of the budget amount by the end of the budget period.

[00:54](javascript:;)In addition to receiving an email, you can use Cloud Pub/Sub notifications to programmatically receive spend updates about this budget.

[01:03](javascript:;)You could even create a Cloud function that listens to the PubSub topic to automate cost management.

[01:10](javascript:;)Here's an example of an email notification.

[01:13](javascript:;)The email contains the project name, the percent of the budget that was exceeded, and the budget amount.

[01:19](javascript:;)Another way to help optimize your GCP spend is to use labels.

[01:24](javascript:;)For example, you could label VM instances that are spread across different regions.

[01:29](javascript:;)Maybe these instances are sending most of their traffic to a different continent, which could incur higher costs.

[01:36](javascript:;)In that case, you might consider relocating some of those instances or using a caching

[01:40](javascript:;)service like Cloud CDN to cache content closer to your users, which reduces your networking spend.

[01:47](javascript:;)I recommend labeling all your resources and exporting your billing data to BigQuery to analyze your spend.

[01:54](javascript:;)BigQuery is Google's scalable, fully managed enterprise data warehouse with SQL and fast response times.

[02:02](javascript:;)Creating a query is as simple as shown in this screenshot, which you will explore in the upcoming lab.

[02:08](javascript:;)You can even visualize spend over time with Data Studio.

[02:12](javascript:;)Data Studio turns your data into informative dashboards and reports that are easy to read, easy to share, and fully customizable.

[02:21](javascript:;)For example, you can slice and dice your billing reports using your labels.

Demo: Billing Administration-

[00:00](javascript:;)In the upcoming lab, you will examine billing data that we exported for you.

[00:05](javascript:;)Let me show you how to export billing data and demonstrate other common activities that a billing administrator performs.

[00:11](javascript:;)These actions cannot be performed in the Qwiklabs environment because of security restrictions.

[00:17](javascript:;)Therefore, I'm going to walk you through them as a demo.

[00:20](javascript:;)So here I am in the GCP console.

[00:23](javascript:;)And what I want to do is navigate to billing.

[00:26](javascript:;)So I'm going to click on the navigation menu and click on Billing.

[00:32](javascript:;)So here I'm provided with an overview.

[00:34](javascript:;)I'm actually using a trial account, as you can see, but the same concepts apply to any account.

[00:41](javascript:;)We can see the consumption for the current month.

[00:44](javascript:;)In my case, I have some promotional credits here.

[00:48](javascript:;)If I had multiple billing accounts, then you'd be able to choose from them right up here.

[00:54](javascript:;)Again, credits, billing account also has a name.

[00:57](javascript:;)If you go to Payment Overview, you'll can see the payment method that is currently selected.

[01:03](javascript:;)Now, the other really big thing we can do is here is set up budgets and alerts.

[01:08](javascript:;)So if I click on that and then click on Create Budget, I first just give it a name.

[01:15](javascript:;)So you My-Budget-Alert.

[01:19](javascript:;)I can select the specific projects that I want these alerts on.

[01:23](javascript:;)It could just be one, it could be several, click Next.

[01:27](javascript:;)Then the type so either specify an exact dollar amount or I start with my last month spend.

[01:33](javascript:;)So if last month I spent would say a certain dollar amount, it would pull that dollar amount directly in there.

[01:39](javascript:;)In our case, let's say we want to target an amount of $500.

[01:45](javascript:;)It's also including credits in this cost.

[01:47](javascript:;)You could disable that.

[01:49](javascript:;)And then we define the thresholds for the alerts.

[01:53](javascript:;)So by default, it sets up this 50, 90 at a 100%, and the dollar amounts are being pulled from the 500 that I just plugged in there earlier.

[02:02](javascript:;)You can also choose between actual and forecasted.

[02:05](javascript:;)And you can read a little bit more what this forecast it is about here.

[02:09](javascript:;)We could remove these.

[02:11](javascript:;)So let's say maybe we wanted to add an earliest threshold already at 25%, then we could do that.

[02:19](javascript:;)We could even go further and actually connect to a cloud pub/sub topic and then do all sorts of automation as we discussed in the slides.

[02:28](javascript:;)So then from here I can just click Finish.

[02:31](javascript:;)And it's then going to be sending me emails around this.

[02:35](javascript:;)I can also see sort of a menu here that shows how far I've gone in the spend.

[02:44](javascript:;)And so I can come in here at any moment and also see, am I close to that 25% mark already?

[02:51](javascript:;)Then the other thing that's pretty interesting on this page is the Transaction page.

[02:55](javascript:;)So this will show all of the different charges.

[02:58](javascript:;)Again, I have credit in here.

[03:01](javascript:;)So every charge is being offset by credit, but you can see all the different usages that I've had across Compute Engine instances, disk space that I've been using.

[03:12](javascript:;)So you could always go in here.

[03:14](javascript:;)Now more interestingly is to probably export all this information.

[03:18](javascript:;)So if I click on Billing Export, I'm presented with two options.

[03:21](javascript:;)I could export to BigQuery or export it as a file, and to enable that I just choose the one I'm interested in.

[03:29](javascript:;)Let's say BigQuery.

[03:30](javascript:;)And then go to edit settings and then I would define where this is going to go to.

[03:36](javascript:;)In this case, I would have to define a BigQuery data set, so you can navigate there and set it up and click Save.

[03:42](javascript:;)And similarly, if I wanted to export to a file, I could edit those settings as well.

[03:48](javascript:;)In this case, could be exported as a CSV or JSON file and stored in a cloud storage bucket.

[03:54](javascript:;)So I can define the name in here.

[03:56](javascript:;)So I would have to create a bucket first and then give it a prefix.

[03:59](javascript:;)And then it's going to export that to there.

[04:02](javascript:;)The other big thing, I can reviews if I click on payment method.

[04:05](javascript:;)I can review the different payment accounts.

[04:08](javascript:;)Payment profiles, payment method if it's a credit card or bank account.

[04:13](javascript:;)You can review all of that information in there.

[04:17](javascript:;)That's how easy it is to administer billing in GCP.

[04:20](javascript:;)A billing administrator can set up accounts and run reports which are ordinary tasks.

[04:25](javascript:;)But becoming familiar with the available options and seeing how these tasks are performed reduces the chances of confusion.

[04:32](javascript:;)For example, you know that reports can be generated in JSON or CSV format.

[04:37](javascript:;)Now more sophisticated processing or filtering of data occurs after the billing is exported, as you will explore in the next lab.

Lab Intro: Examining Billing Data with BigQuery-

[00:00](javascript:;)Let's examine billing data with BigQuery.

[00:03](javascript:;)In this lab, you will sign into BigQuery and create a dataset.

[00:07](javascript:;)In this dataset, you will create a table by importing billing data that is stored in a Cloud Storage bucket.

[00:13](javascript:;)Next, you will run simple queries on the imported data, and then you will run more complex queries on a larger dataset.

Examining Billing Data with BigQuery- Lab to be done

Lab Review: Examining Billing Data with BigQuery-

[00:00](javascript:;)In this lab, you imported billing data into BigQuery that had been exported as a CSV file.

[00:05](javascript:;)You first ran a simple query on that data, next you accessed a shared dataset containing more than 22,000 records of billing information, you

[00:14](javascript:;)then ran a variety of queries on that data to explore how you can use BigQuery to gain insight into your resources billing consumption.

[00:24](javascript:;)If you use BigQuery on a regular basis, you'll start to develop your own queries for searching out where resources are being consumed in your application.

[00:32](javascript:;)You can also monitor changes in resource consumption over time.

[00:36](javascript:;)This kind of analysis is an input to capacity planning and can help you determine

[00:41](javascript:;)how to scale up your application to meet growth or scale down your application for efficiency.

[00:48](javascript:;)Welcome to the walk-through of the lab examining billing data with BigQuery.

[00:53](javascript:;)At this point in the lab, I have logged in with the username and password that Qwiklabs has provided me from the lab.

[01:00](javascript:;)So the first task is to use BigQuery to import data.

[01:05](javascript:;)So what I did is as the Billing Administrator, I exported my billing data and put it in a bucket.

[01:13](javascript:;)So I am going to go into BigQuery and I'm going to import some stuff.

[01:19](javascript:;)So BigQuery?

[01:23](javascript:;)Yes.

[01:23](javascript:;)The Cloud Council.

[01:25](javascript:;)Thank you.

[01:29](javascript:;)Make sure that you are logged in to BigQuery and have the correct Qwiklabs project ID selected at the top.

[01:38](javascript:;)So I'm going to go here and I'm going to click Create dataset and I am going to call it imported billing data.

[01:51](javascript:;)My data location is in the US and I want it to expire one day afterwards.

[02:01](javascript:;)I'm going to hit Create dataset.

[02:06](javascript:;)You can see my dataset is created and I should see it right here.

[02:12](javascript:;)There it is.

[02:12](javascript:;)So now I'm going to create a table in that dataset.

[02:18](javascript:;)Table.

[02:18](javascript:;)For the source, I'm going to use Cloud Storage, I'm going to copy and paste the bucket location from the lab and it is in CSV format.

[02:34](javascript:;)For destination, I'm using that and native table and I am going to call it sampleinfotable.

[02:46](javascript:;)Under schema, I I'm going to hit auto detect so it detects a schema and input parameters from the dataset.

[02:55](javascript:;)I'm going to open up the advance and I am going to specify that I want

[03:02](javascript:;)to skip one row because that's the headers, and then I'm going to hit Create table.

[03:10](javascript:;)So this is a point where you can check the progress in your lab.

[03:13](javascript:;)If you click check my progress, it's going to check that you have dataset at a table

[03:18](javascript:;)and that you have imported that data into that table, and you should get five points for that.

[03:23](javascript:;)So Task 2, you are going to examine the data that you've just input.

[03:27](javascript:;)So I'm going to click on my table and it's going to, by default, show the schema.

[03:35](javascript:;)I can click Details and it's going to tell me a little bit more information about number of rows.

[03:41](javascript:;)You can see it has 44 rows, it's a pretty small table.

[03:45](javascript:;)I can hit Preview and it's going to show me a couple of the first rows of the table.

[03:51](javascript:;)So it now wants me in the lab.

[03:55](javascript:;)There are some formative questions that are going to ask you just to make sure that you're understanding the learning.

[04:02](javascript:;)So I'm not going to go over that in the walk-through because that is more and to make sure that you're understanding what you're doing.

[04:08](javascript:;)So I'm going to go to Task 3 where we're composing a simple query.

[04:12](javascript:;)A couple of cool things about BigQuery, if you are in a table by default, if you click Query table, it's going

[04:22](javascript:;)to auto-populate the query editor with the project dataset table for you and then you just specify what you want to look at.

[04:35](javascript:;)So I do select star and I'm doing this.

[04:41](javascript:;)Oops.

[04:41](javascript:;)Where cost is greater than zero.

[04:47](javascript:;)So I just want to see in this table how much of it.

[04:52](javascript:;)I only want to see the rows where the cost is more than zero.

[04:56](javascript:;)You can see right here it's validating that my SQL or my SQL is right.

[05:03](javascript:;)I'm going to hit Run.

[05:06](javascript:;)Here are my query results and you can see out of a table that has 44,

[05:11](javascript:;)there are 20 rows in this table that actually have costs that are more than zero.

[05:17](javascript:;)So again, there are a couple more questions that you can answer and you can also check your progress that you've run this query.

[05:26](javascript:;)So if you run the query, then it's going to give you another five points.

[05:31](javascript:;)At this point, you're actually done with the points that are awarded in the lab, but you still have another task to go into a little more complex query.

[05:40](javascript:;)So I'm going to go ahead and copy the query from Task 4, and I'm going to erase this and paste it in.

[05:49](javascript:;)It's a valid query here.

[05:51](javascript:;)I'm going to hit Run, then I'm going to verify that the result has what my lab is telling me is, supposed to return 22,537 lines of billing data.

[06:04](javascript:;)I can see right here, that is correct.

[06:08](javascript:;)Let's say I wanted to find the latest 100 records where the charges were greater than zero.

[06:14](javascript:;)So I'm going to copy paste the query that's provided to me, and make sure it's valid.

[06:20](javascript:;)Always good to check that your SQL is valid.

[06:23](javascript:;)Hit Run, and it is going to show me the last 100 records where charges were greater than zero.

[06:33](javascript:;)Let's say I wanted to find all of the charges that were more than $3, the next query shows you that.

[06:39](javascript:;)You can feel free to click through each one of these more complex queries and feel free to try out some queries of your own.

[06:47](javascript:;)If you wanted to just peruse the data and figure out maybe the last two days of billing anything that was over $10.

[06:55](javascript:;)Any kind of question that you might need to provide data for to your senior leadership about your billing of resource usage in GCP.

[07:06](javascript:;)After all of these complex queries, in review, you imported billing data that was technically exported for you from

[07:15](javascript:;)a billing admin into BigQuery, and then you run a simple query and then you ran some more complex queries.

[07:23](javascript:;)I hope you enjoyed the lab.

[07:24](javascript:;)Thank you.

Quiz: Resource Management-

How do quotas protect Google Cloud customers?

By preventing uncontrolled consumption of resources.

Quotas are established at reasonable defaults for common cloud usage and proof of concept activities. If you are planning to scale up a production cloud solution you may need to request that the quotas be raised. This is a reasonable checkpoint to verify that actions that might result in a large consumption of resources are reviewed.

A budget is set at $500 and an alert is set at 100%. What happens when the full amount is used?   
A notification email is sent to the Billing Administrator.

Budgets in Google Cloud are not a way to prevent spending or stop resources. They are a tool for raising awareness about the consumption of resources so that a business can implement its own consumption management processes.

No resources in Google Cloud can be used without being associated with...

A project.

All resources in Google Cloud are tracked and their consumption is logged against a project. A project relates resources to a billing method.

Module Review-

[00:00](javascript:;)Person: In this module, we covered the Cloud Resource Manager and went into quotas, labels, and billing.

[00:06](javascript:;)Then we analyzed the billing data with BigQuery in a lab.

[00:11](javascript:;)Reporting is an important part of resource management.

[00:15](javascript:;)You can generate reports to track consumption and to establish accountability.

[00:19](javascript:;)A key principle in Google Cloud is transparency, and that means it's straightforward to access and process consumption data, as you observed in this module.

RESOURCE MONITORING:

Module Overview-

[00:00](javascript:;)In this module, I’ll give you an overview of the resource monitoring options in Google Cloud.

[00:05](javascript:;)The features covered in this module rely on Google Cloud’s operations suite, a service that provides monitoring, logging, and diagnostics for your applications.

[00:16](javascript:;)In this module we are going to explore the Cloud Monitoring, Cloud Logging, Error Reporting, Cloud Trace, and Cloud Profiler services.

[00:24](javascript:;)Let me start by giving you a high-level overview of Google Cloud’s operations suite and its features.

[00:29](javascript:;)Let me start by giving you a high-level overview of Google Cloud’s operations suite and its features.

Stackdriver Overview-

[00:00](javascript:;)Google Cloud’s operations suite dynamically discovers cloud resources and application services based on deep integration with Google Cloud and Amazon Web Services.

[00:10](javascript:;)Because of its smart defaults, you can have core visibility into your cloud platform in minutes.

[00:17](javascript:;)This provides you with access to powerful data and analytics tools plus collaboration with many different third-party software providers.

[00:26](javascript:;)As mentioned earlier, Google Cloud’s operations suite has services for monitoring, logging, error reporting, and fault tracing.

[00:36](javascript:;)You only pay for what you use, and there are free usage allotments so that you can get started with no upfront fees or commitments.

[00:44](javascript:;)For more information about pricing, refer to the link in the Course Resources.

[00:48](javascript:;)Now, in most other environments, these services are handled by completely different packages, or by a loosely integrated collection of software.

[00:57](javascript:;)When you see these functions working together in a single, comprehensive, and integrated service, you'll realize how important that is to creating reliable, stable, and maintainable applications.

[01:11](javascript:;)Google Cloud’s operations suite also supports a rich and growing ecosystem of technology partners, as shown on this slide.

[01:19](javascript:;)This helps expand the IT ops, security, and compliance capabilities available to Google Cloud customers.

[01:27](javascript:;)For more information about integrations, refer to the link in the Course Resources.

Monitoring-

[00:00](javascript:;)Now that you understand Google Cloud’s operations suite from a high-level perspective, let’s look at Cloud Monitoring.

[00:07](javascript:;)Monitoring is important to Google because it is at the base of site reliability engineering, or SRE.

[00:13](javascript:;)SRE is a discipline that applies aspects of software engineering to operations whose goals are to create ultra-scalable and highly reliable software systems.

[00:25](javascript:;)This discipline has enabled Google to build, deploy, monitor, and maintain some of the largest software systems in the world.

[00:34](javascript:;)Cloud Monitoring dynamically configures monitoring after resources are deployed and has intelligent defaults that allow you to easily create charts for basic monitoring activities.

[00:44](javascript:;)This allows you to monitor your platform, system, and application metrics by ingesting data, such as metrics, events, and metadata.

[00:53](javascript:;)You can then generate insights from this data through dashboards, charts, and alerts.

[00:58](javascript:;)For example, you can configure and measure uptime and health checks that send alerts via email.

[01:04](javascript:;)A metrics scope is the root entity that holds monitoring and configuration information in Cloud Monitoring.

[01:12](javascript:;)Each metrics scope can have between 1 and 100 monitored projects.

[01:17](javascript:;)You can have as many metrics scopes as you want, but Google Cloud projects and AWS accounts can't be monitored by more than one metrics scope.

[01:26](javascript:;)A metrics scope contains the custom dashboards, alerting policies, uptime checks, notification channels, and group definitions that you use with your monitored projects.

[01:37](javascript:;)A metrics scope can access metric data from its monitored projects, but the metrics data and log entries remain in the individual projects.

[01:45](javascript:;)The first monitored Google Cloud project in a metrics scope is called the hosting project, and it must be specified when you create the metrics scope.

[01:55](javascript:;)The name of that project becomes the name of your metrics scope.

[01:58](javascript:;)To access an AWS account, you must configure a project in Google Cloud to hold the AWS Connector.

[02:05](javascript:;)Because metrics scopes can monitor all your Google Cloud projects in a single place, a metrics scope is

[02:10](javascript:;)a “single pane of glass” through which you can view resources from multiple Google Cloud projects and AWS accounts.

[02:19](javascript:;)All users of Google Cloud’s operations suite with access to that metrics scope have access to all data by default.

[02:26](javascript:;)This means that a role assigned to one person on one project applies equally to all projects monitored by that metrics scope.

[02:34](javascript:;)In order to give people different roles per-project and to control visibility to data, consider placing the monitoring of those projects in separate metrics scopes.

[02:44](javascript:;)Cloud Monitoring allows you to create custom dashboards that contain charts of the metrics that you want to monitor.

[02:50](javascript:;)For example, you can create charts that display your instances’ CPU utilization, the packets or bytes sent

[02:56](javascript:;)and received by those instances, and the packets or bytes dropped by the firewall of those instances.

[03:02](javascript:;)In other words, charts provide visibility into the utilization and network traffic of your VM instances, as shown on this slide.

[03:11](javascript:;)These charts can be customized with filters to remove noise, groups to reduce the number of time series, and aggregates to group multiple time series together.

[03:22](javascript:;)For a full list of supported metrics, please refer to the documentation.

[03:26](javascript:;)Now, although charts are extremely useful, they can only provide insight while someone is looking at them.

[03:33](javascript:;)But what if your server goes down in the middle of the night or over the weekend?

[03:37](javascript:;)Do you expect someone to always look at dashboards to determine whether your servers are available or have enough capacity or bandwidth?

[03:44](javascript:;)If not, you want to create alerting policies that notify you when specific conditions are met.

[03:51](javascript:;)For example, as shown on this slide, you can create an alerting policy when the network egress of your VM instance goes above a certain threshold for a specific timeframe.

[04:02](javascript:;)When this condition is met, you or someone else can be automatically notified through email, SMS, or other channels in order to troubleshoot this issue.

[04:10](javascript:;)You can also create an alerting policy that monitors your usage of Google Cloud’s operations suite and alerts you when you approach the threshold for billing.

[04:19](javascript:;)For more information about this, please refer to the documentation.

[04:21](javascript:;)Here is an example of what creating an alerting policy looks like.

[04:24](javascript:;)On the left, you can see an HTTP check condition on the summer01 instance.

[04:30](javascript:;)This will send an email that is customized with the content of the documentation section on the right.

[04:35](javascript:;)Let’s discuss some best practices when creating alerts: \* We recommend alerting on symptoms, and not necessarily causes.

[04:44](javascript:;)For example, you want to monitor failing queries of a database and then identify whether the database is down.

[04:51](javascript:;)\* Next, make sure that you are using multiple notification channels, like email and SMS.

[04:56](javascript:;)This helps avoid a single point of failure in your alerting strategy.

[05:02](javascript:;)\* We also recommend customizing your alerts to the audience’s needs by describing what actions need to be taken or what resources need to be examined.

[05:10](javascript:;)\* Finally, avoid noise, because this will cause alerts to be dismissed over time.

[05:16](javascript:;)Specifically, adjust monitoring alerts so that they are actionable and don’t just set up alerts on everything possible.

[05:24](javascript:;)Uptime checks can be configured to test the availability of your public services from locations around the world, as you can see on this slide.

[05:31](javascript:;)The type of uptime check can be set to HTTP, HTTPS, or TCP.

[05:38](javascript:;)The resource to be checked can be an App Engine application, a Compute Engine instance, a URL of a host, or an AWS instance or load balancer.

[05:47](javascript:;)For each uptime check, you can create an alerting policy and view the latency of each global location.

[05:54](javascript:;)Here is an example of an HTTP uptime check.

[05:59](javascript:;)The resource is checked every minute with a 10-second timeout.

[06:03](javascript:;)Uptime checks that do not get a response within this timeout period are considered failures.

[06:08](javascript:;)So far there is a 100% uptime with no outages.

[06:11](javascript:;)If the standard metrics provided by Cloud Monitoring do not fit your needs, you can create custom metrics.

[06:20](javascript:;)For example, imagine a game server that has a capacity of 50 users.

[06:24](javascript:;)What metric indicator might you use to trigger scaling events?

[06:29](javascript:;)From an infrastructure perspective, you might consider using CPU load or perhaps network traffic load as values that are somewhat correlated with the number of users.

[06:39](javascript:;)But with a Custom Metric, you could actually pass the current number of users directly from your application into Cloud Monitoring.

[06:44](javascript:;)To get started with creating custom metrics, please refer to the documentation.

Lab Intro: Resource Monitoring-

[00:00](javascript:;)Let's take some of the monitoring concepts that we just discussed and apply them in a lab.

[00:05](javascript:;)In this lab, you will learn how to use Stackdriver monitoring to gain insight into applications that run on GCP.

[00:12](javascript:;)Specifically, you will enable Stackdriver monitoring, add charts to dashboards, and create alerts, resource groups, and up-time checks.

Resource Monitoring- Lab to be done

Lab Review: Resource Monitoring-

[00:00](javascript:;)In this lab you got an overview of stackdriver monitoring.

[00:04](javascript:;)You learned how to monitor your project, create alerts with multiple conditions, add charts to dashboards, create resource groups, and create uptime checks for your services.

[00:14](javascript:;)Monitoring is critical to your applications health.

[00:17](javascript:;)And Stackdriver provides a rich set of features for monitoring your infrastructure, visualizing the monitoring data and triggering alerts and events for you.

[00:26](javascript:;)You can stay for a lab walkthrough, but remember that GCP is user interface can change, so your environment might look slightly different.

[00:34](javascript:;)Welcome to the lab walkthrough for resource monitoring with stackdriver.

[00:39](javascript:;)At this point I have logged into the GCP console with the credentials that the Quick Labs lab has provided me.

[00:47](javascript:;)And in the first task I am going to verify that the proper Vms had been set up for me using deployment manager in this lab.

[00:56](javascript:;)And as you can see, there are three Vms.

[01:00](javascript:;)3 engine X stacks right here.

[01:04](javascript:;)So now that I verified that my instances are running and were created for me, I am going to go to stackdriver monitoring which will open in a new tab.

[01:15](javascript:;)And then it is going to set up the workspace for my project.

[01:26](javascript:;)And this can take a few minutes.

[01:28](javascript:;)So just be patient or go get a cup of coffee and come back.

[01:33](javascript:;)Once your workspace has been set up for you, you are going to be redirected to the monitoring overview page.

[01:43](javascript:;)There are some questions in the labs that are going to ask you some questions and those are just making sure that your understanding and actually reading.

[01:50](javascript:;)But you don't have to fill those out in order to get the full score for the labs.

[01:55](javascript:;)So in task two, we're going to create a dashboard, so I'm going to go here.

[02:02](javascript:;)To monitoring overview when I click create dashboard.

[02:09](javascript:;)I am going to name it my dashboard instead of untitled hit enter.

[02:16](javascript:;)And then I'm going to add a chart.

[02:20](javascript:;)For the title, I am going to say this is my chart.

[02:25](javascript:;)And I am going to find GCE.

[02:30](javascript:;)VM instance.

[02:33](javascript:;)For metrics, I am going to select CPU utilization.

[02:39](javascript:;)CPU utilization and for filter.

[02:50](javascript:;)Where is filter here?

[02:51](javascript:;)I'm going to add a filter.

[02:54](javascript:;)There are various options you can filter by resource label by metadata label.

[03:01](javascript:;)I'm not going to add any filters, I want to see everything.

[03:06](javascript:;)And then we click here on view options.

[03:10](javascript:;)There's a couple of chart modes.

[03:12](javascript:;)There's color mode X Ray mode.

[03:15](javascript:;)You can preview it on the right.

[03:18](javascript:;)Stats mode and like that.

[03:22](javascript:;)So I actually like the X Ray mode, so I'm going to go ahead and click that.

[03:26](javascript:;)And then you're going to hit save to add the chart to your dashboard.

[03:29](javascript:;)There it is, looking nice.

[03:33](javascript:;)So then we also have a metrics explorer which allows you to examine resources and metrics without having to create a chart on the dashboard.

[03:41](javascript:;)So if I go to resources metrics explorer.

[03:46](javascript:;)Find resource type in metric, I can type any metric or resource name.

[03:51](javascript:;)So let's say I do CPU utilization.

[03:55](javascript:;)And as you can see I didn't have to add this, but I could still explore it.

[04:00](javascript:;)Again, you'll have another question in the lab, but that is to prompt your understanding.

[04:05](javascript:;)So now, I'm going to create an alert and add the first condition.

[04:09](javascript:;)So I'm going to go here and create a policy.

[04:15](javascript:;)And I'm going to click add condition.

[04:18](javascript:;)Here I'm going to do GCE VM instance.

[04:23](javascript:;)And for metrics, I'm going to use CPU usage.

[04:30](javascript:;)For condition, I'm going to say is above threshold.

[04:39](javascript:;)Of one minute, the threshold is 20.

[04:46](javascript:;)And then I'm going to hit save.

[04:49](javascript:;)And I'm going to add another condition.

[04:54](javascript:;)And then what said do I do it for another VM?

[05:03](javascript:;)So if I do this one.

[05:06](javascript:;)Maybe I do it for another metric going to, do you?

[05:11](javascript:;)Stu reserved course.

[05:15](javascript:;)And then above 15.

[05:19](javascript:;)I'm going to head, save.

[05:24](javascript:;)So now in policy triggers I'm going to trigger when all conditions are met.

[05:32](javascript:;)Then I'm going to configure the notifications so that I can be actually told that this has triggered.

[05:39](javascript:;)And I'm going to click here.

[05:43](javascript:;)Email, and I'm going to add some email going to do fake email.

[05:49](javascript:;)I want you guys spamming me.

[05:53](javascript:;)And then add.

[05:59](javascript:;)That has been added.

[06:02](javascript:;)And then I'm going to stick, skip the documentation step, but in reality this is a pretty important part of your notification.

[06:09](javascript:;)You want to say.

[06:11](javascript:;)What happened, why whoever it is that's getting notified is being alerted and a best practices to actually tell them how to maybe fix it.

[06:20](javascript:;)because otherwise that's not a very useful notification if they don't know how they can fix it.

[06:26](javascript:;)And then you're going to name it, and this is my.

[06:30](javascript:;)I first alerting policy.

[06:34](javascript:;)And then I'm going to hit save.

[06:36](javascript:;)This is a checkpoint in the lab where you can check your progress that you have created an alerting policy.

[06:42](javascript:;)The next task you are going to create some groups here.

[06:47](javascript:;)Create group, I'm going to give it a name.

[06:51](javascript:;)VM instances.

[06:54](javascript:;)Name, I am going to select.

[07:02](javascript:;)Contains and I'm going to type engine X. And I'm going to save group.

[07:14](javascript:;)And you can see it's showing me instances.

[07:18](javascript:;)All three of my instances because the name matches Anjanette stack.

[07:24](javascript:;)And again, you're going to have another.

[07:26](javascript:;)Question to make sure that your understanding and reading all of the extra tidbits that are available in the lab.

[07:34](javascript:;)So now, we're going to go back to the Dashboard.

[07:39](javascript:;)You' re going to go to uptime checks.

[07:43](javascript:;)Overview and then we're going to add an uptight and check.

[07:47](javascript:;)So in here we're going to add a title, so my first uptime check were using HTTP.

[07:56](javascript:;)It is to check an instance in, applies to a group and I'm going to

[08:01](javascript:;)select group I created which is VM instances and I'm going to check every one minute.

[08:07](javascript:;)I'm going to hit save.

[08:10](javascript:;)You could also hit test for it and make sure that it works.

[08:13](javascript:;)I'm going to say no thanks.

[08:14](javascript:;)I don't want to create the alert policy right now.

[08:16](javascript:;)So this is the last piece where you can hit check my progress.

[08:19](javascript:;)It's going to make sure that you created that uptime check, and if so, you're going to get the full points for the lab.

[08:25](javascript:;)So in this lab, we got to walk through monitoring your projects, creating a stackdriver workspace which gets created for you.

[08:31](javascript:;)Creating some alerts with multiple condition, adding some charts to a dashboard, and creating resource groups, and finally we created an uptime check for your services.

[08:41](javascript:;)Hope you enjoyed it.

Logging-

[00:00](javascript:;)Monitoring is the basis of Google Cloud’s operation suite, but the service also provides logging, error reporting, and tracing.

[00:08](javascript:;)Let’s learn about logging.

[00:12](javascript:;)Cloud Logging allows you to store, search, analyze, monitor, and alert on log data and events from Google Cloud and AWS.

[00:24](javascript:;)It is a fully managed service that performs at scale and can ingest application and system log data from thousands of VMs.

[00:34](javascript:;)Logging includes storage for logs, a user interface called Logs Explorer, and an API to manage logs programmatically.

[00:41](javascript:;)The service lets you read and write log entries, search and filter your logs, and create log-based metrics.

[00:51](javascript:;)Logs are only retained for 30 days, but you can export your logs to Cloud Storage buckets, BigQuery datasets, and Pub/Sub topics.

[01:00](javascript:;)Exporting logs to Cloud Storage makes sense for storing logs for more than 30 days, but why should you export to BigQuery or Pub/Sub?

[01:11](javascript:;)Exporting logs to BigQuery allows you to analyze logs and even visualize them in Looker Studio.

[01:17](javascript:;)BigQuery runs extremely fast SQL queries on gigabytes to petabytes of data.

[01:24](javascript:;)This allows you to analyze logs, such as your network traffic, so that you can better understand

[01:29](javascript:;)traffic growth to forecast capacity, network usage to optimize network traffic expenses, or network forensics to analyze incidents.

[01:38](javascript:;)For example, in this screenshot I queried my logs to identify the top IP addresses that have exchanged traffic with my web server.

[01:48](javascript:;)Depending on where these IP addresses are, and who they belong to, I could relocate part of my infrastructure to

[01:55](javascript:;)save on networking costs or deny some of these IP addresses if I don’t want them to access my web server.

[02:03](javascript:;)If you want to visualize your logs, I recommend connecting your BigQuery tables to Looker Studio.

[02:09](javascript:;)Looker Studio transforms your raw data into the metrics and dimensions that you can use to create easy-to-understand reports and dashboards.

[02:19](javascript:;)I mentioned that you can also export logs to Pub/Sub.

[02:23](javascript:;)This enables you to stream logs to applications or endpoints.

Error Reporting-

[00:00](javascript:;)Let’s learn about another feature of Google Cloud’s operations suite: Error Reporting.

[00:05](javascript:;)Error Reporting counts, analyzes, and aggregates the errors in your running cloud services.

[00:12](javascript:;)A centralized error management interface displays the results with sorting and filtering capabilities, and you can even set up real-time notifications when new errors are detected.

[00:24](javascript:;)Currently, Error Reporting is generally available for App Engine on both standard and flexible environments, Apps Script, Compute Engine, Cloud Functions, Cloud Run, Google Kubernetes Engine, and Amazon EC2.

[00:41](javascript:;)In terms of programming languages, the exception stack trace parser is able to process Go, Java, .

[00:48](javascript:;)NET, Node.js, PHP, Python, and Ruby.

Tracing-

[00:00](javascript:;)Tracing is another Cloud Operations feature integrated into Google Cloud.

[00:05](javascript:;)Cloud Trace is a distributed tracing system that collects latency data from your applications and displays it in the Google Cloud console.

[00:14](javascript:;)You can track how requests propagate through your application and receive detailed near real-time performance insights.

[00:23](javascript:;)Cloud Trace automatically analyzes all of your application's traces to generate in-depth latency reports that surface performance degradations

[00:32](javascript:;)and can capture traces from App Engine, HTTP(S) load balancers, and applications instrumented with the Cloud Trace API.

[00:43](javascript:;)Managing the amount of time it takes for your application to handle incoming requests and perform operations is an important part of managing overall application performance.

[00:55](javascript:;)Cloud Trace is actually based on the tools used at Google to keep our services running at extreme scale.

Profiling-

[00:00](javascript:;)Finally, let’s cover the last feature of Google Cloud’s operations suite in this module, which is the profiler.

[00:08](javascript:;)Poorly performing code increases the latency and cost of applications and web services every day.

[00:16](javascript:;)Cloud Profiler continuously analyzes the performance of CPU or memory-intensive functions executed across an application.

[00:25](javascript:;)While it’s possible to measure code performance in development environments, the results generally don’t map well to what’s happening in production.

[00:35](javascript:;)Many production profiling techniques either slow down code execution or can only inspect a small subset of a codebase.

[00:43](javascript:;)Profiler uses statistical techniques and extremely low-impact instrumentation that runs across all production application instances to provide a complete picture of an application’s performance without slowing it down.

[00:58](javascript:;)Profiler allows developers to analyze applications running anywhere, including Google Cloud, other cloud platforms, or on-premises, with support for Java, Go, Node.js, and Python.

Error Reporting and Debugging- Lab to be done

Quiz: Resource Monitoring-

What is the purpose of the Cloud Trace service?

Reporting on latency as part of managing performance.

Cloud Trace provides latency sampling and reporting for Google App Engine, Google HTTP(S) load balancers, and applications instrumented with the Cloud Trace SDKs. Reporting includes per-URL statistics and latency distributions.

Google Cloud’s operations suite integrates several technologies, including monitoring, logging, error reporting, and debugging that are commonly implemented in other environments as separate solutions using separate products. What are key benefits of integration of these services?

Reduces overhead, reduces noise, streamlines use, and fixes problems faster

Integration with Google Cloud’s operations suite streamlines and unifies these traditionally independent services, making it much easier to establish procedures around them and to use them in continuous ways.

What is the foundational process at the base of Google's Site Reliability Engineering (SRE)?

Monitoring

Before you can take any of the other actions, you must first be monitoring the system.

Module Review-

[00:00](javascript:;)In this module, I gave you an overview of Google Cloud’s operations suite and its monitoring, logging, error reporting, fault tracing, and profiling features.

[00:11](javascript:;)Having all of these integrated into Google Cloud allows you to operate and maintain your applications, which is known as site reliability engineering or SRE.

[00:23](javascript:;)If you’re interested in learning more about SRE, you can explore the SRE book or some of our SRE courses.

COURSE REVIEW-

[00:00](javascript:;)Thank you for taking the “Essential Cloud Infrastructure: Core Services” course.

[00:04](javascript:;)I hope you have a better understanding of how to administer IAM, choose between the different data storage services in GCP, examine billing of GCP resources and monitor those resources.

[00:17](javascript:;)Hopefully the demos and labs made you feel more comfortable using the different GCP services that we covered.

[00:24](javascript:;)Next, I recommend enrolling in the “Elastic Cloud Infrastructure: Scaling and Automation” course, of the “Architecting with Google Compute Engine” series.

[00:33](javascript:;)In that course, we start by going over the different options to interconnect networks to enable you to connect your infrastructure to GCP.

[00:43](javascript:;)Next, we’ll go over GCPs load balancing and autoscaling services, which you will get to explore directly.

[00:50](javascript:;)Then, we’ll cover infrastructure automation services like Terraform, So that you can automate the deployment of GCP

[00:57](javascript:;)infrastructure services Lastly, we'll talk about other managed services that you might want to leverage in GCP.