

LIVRO: Google Cloud Certified Associate Cloud Engineer Study Guide

Edição: Segunda edição

Ano: 2023

Autor: Dan Sullivan

Introdução

O Google Cloud é um dos principais serviços de nuvem pública que fornece a seus usuários parte do mesmo software, hardware e infraestrutura de rede usados para alimentar os serviços do Google. Empresas, organizações e indivíduos podem lançar servidores em minutos, armazenar petabytes de dados e implementar nuvens virtuais globais com o Google Cloud. Ele inclui uma interface de console fácil de usar, ferramentas de linha de comando e interfaces de programação de aplicações (APIs) para gerenciar recursos na nuvem. Os usuários podem trabalhar com recursos gerais, como máquinas virtuais (VMs) e discos persistentes, ou optar por serviços altamente focados para Internet das Coisas (IoT), aprendizado de máquina, mídia e outros domínios especializados.

Implantar e gerenciar aplicações e serviços no Google Cloud requer um entendimento claro de como o Google estrutura contas de usuários e gerencia identidades e controles de acesso; você também precisa entender as vantagens e desvantagens de usar vários serviços. Engenheiros Associados Certificados em Cloud demonstraram o conhecimento e as habilidades necessárias para implantar e operar infraestrutura, serviços e redes no Google Cloud.

Este guia de estudo é projetado para ajudá-lo a entender o Google Cloud em profundidade para que você possa atender às necessidades daqueles que operam recursos no Google Cloud. Sim, este livro vai, claro, ajudar você a passar no exame de certificação de Engenheiro Associado em Cloud, mas não é um guia de crônica para exames. Você aprenderá mais do que é necessário para passar no exame; você entenderá como atender aos desafios diários enfrentados por engenheiros de nuvem, incluindo a escolha de serviços, gerenciamento de usuários, implantação e monitoramento de infraestrutura, e ajudando a mapear requisitos de negócios em soluções baseadas em nuvem.

Cada capítulo deste livro cobre um único tópico e inclui uma seção “Essenciais para o Exame” que destaca as informações chave que você deve saber para passar no exame de certificação. Também há exercícios para ajudá-lo a revisar e reforçar seu entendimento sobre o tópico do capítulo. Questões de exemplo são incluídas no final de cada capítulo para que você possa ter uma ideia dos tipos de questões que verá no exame. O livro também inclui flashcards e exames práticos que cobrem todos os tópicos que você aprenderá com este guia.

O Que Este Livro Aborda?

Este livro descreve produtos e serviços no Google Cloud. Não inclui tópicos de administração do G Suite.

Capítulo 1: Visão Geral do Google Cloud Platform

No capítulo inicial, investigamos os tipos de serviços fornecidos pelo Google Cloud, que incluem serviços de computação, armazenamento e rede, bem como serviços especializados, como produtos de aprendizado de máquina. Este capítulo também descreve algumas das principais diferenças entre a computação em nuvem e a computação em data center ou local.

Capítulo 2: Serviços de Computação em Nuvem do Google

Este capítulo fornece uma visão geral dos serviços de infraestrutura, como computação, armazenamento e rede. Introduz o conceito de gerenciamento de identidade e serviços relacionados. Também apresenta tópicos de DevOps e ferramentas para implantação e monitoramento de aplicações e recursos. O Google Cloud inclui uma lista crescente de serviços especializados, como serviços de aprendizado de máquina e processamento de linguagem natural. Estes são brevemente discutidos neste capítulo. O capítulo apresenta a estrutura organizacional do Google Cloud, com um olhar para regiões e zonas.

Capítulo 3: Projetos, Contas de Serviço e Cobrança

Uma das primeiras coisas que você fará ao começar a trabalhar com o Google Cloud é configurar suas contas. Neste capítulo, você aprenderá como recursos em contas são organizados em organizações, pastas e projetos. Você aprenderá como criar e editar essas estruturas. Você também verá como habilitar APIs para projetos específicos, bem como gerenciar identidades de usuários e seus controles de acesso. Este capítulo descreve como criar contas de cobrança e vinculá-las a projetos. Você também aprenderá a criar orçamentos e definir alertas de cobrança para ajudá-lo a gerenciar custos.

Capítulo 4: Introdução à Computação no Google Cloud

Neste capítulo, você verá a variedade de opções disponíveis para executar aplicações e serviços no Google Cloud. As opções incluem o Compute Engine, que fornece VMs executando sistemas operacionais Linux ou Windows. Cloud Run e App Engine são opções de plataforma como serviço (PaaS) que permitem aos desenvolvedores executar suas aplicações sem precisarem se preocupar em gerenciar VMs. Se você estiver executando múltiplas aplicações e serviços, pode querer tirar vantagem de contêineres, que são uma alternativa leve às VMs. Você aprenderá sobre contêineres e como gerenciá-los com o Kubernetes Engine. Este capítulo também introduz o Cloud Functions, que é para tarefas de curta duração e orientadas a eventos, como acionar o processamento de uma imagem carregada no Cloud Storage.

Capítulo 5: Computação com Máquinas Virtuais do Compute Engine

Neste capítulo, você aprenderá como configurar VMs, incluindo a seleção de CPU, memória, opções de armazenamento e imagens do sistema operacional. Você aprenderá como usar o Google Cloud Console e o Cloud Shell para trabalhar com VMs. Além disso, você verá como instalar a interface de linha de comando e o SDK, que você

usará para iniciar e parar VMs. O capítulo também descreve como habilitar o acesso à rede para VMs.

Capítulo 6: Gerenciando Máquinas Virtuais

No capítulo anterior, você aprendeu a criar VMs, e neste capítulo você aprenderá como gerenciar VMs individuais e grupos de VMs. Você começará gerenciando uma única instância de uma VM usando o console do Google Cloud e, em seguida, realizará as mesmas operações usando o Cloud Shell e a linha de comando. Você também aprenderá a visualizar VMs que estão sendo executadas atualmente. Em seguida, você aprenderá sobre grupos de instâncias, que permitem criar conjuntos de VMs que você pode gerenciar como uma unidade única. Na seção sobre grupos de instâncias, você aprenderá a diferença entre grupos de instâncias gerenciadas e não gerenciadas. Você também aprenderá sobre instâncias preemptivas, que são VMs de baixo custo que podem ser desligadas pelo Google. Você aprenderá sobre as compensações de custo-benefício das instâncias preemptivas. Finalmente, o capítulo se encerra com diretrizes para gerenciar VMs.

Capítulo 7: Computação com Kubernetes

Este capítulo introduz o Kubernetes Engine, o serviço gerenciado de Kubernetes do Google. Kubernetes é uma plataforma de orquestração de contêineres criada e liberada como código aberto pelo Google. Neste capítulo, você aprenderá os conceitos básicos de contêineres, orquestração de contêineres e a arquitetura do Kubernetes. A discussão incluirá uma visão geral dos objetos do Kubernetes, como pods, serviços, volumes e namespaces, bem como controladores do Kubernetes, como ReplicaSets, Deployments e Jobs.

Em seguida, o capítulo passa para a implantação de um cluster Kubernetes usando o console do Google Cloud, Cloud Shell e SDK. Você também verá como implantar pods, que inclui baixar uma imagem Docker existente, construir uma imagem Docker, criar um pod e então implantar uma aplicação no cluster Kubernetes. Claro, você precisará saber como monitorar um cluster de servidores. Este capítulo fornece uma descrição de como configurar monitoramento e registro com o Cloud Operations, que é o serviço de monitoramento de aplicações, serviços, contêineres e infraestrutura do Google.

Capítulo 8: Gerenciando Clusters Kubernetes no Modo Padrão

Neste capítulo, você aprenderá os conceitos básicos de gerenciamento de um cluster Kubernetes, incluindo visualizar o status do cluster, visualizar o conteúdo do repositório de imagens, ver detalhes sobre imagens no repositório e adicionar, modificar e remover nós, pods e serviços. Assim como no capítulo sobre gerenciamento de VMs, neste capítulo você aprenderá como realizar operações de gerenciamento com as três ferramentas de gerenciamento: console do Google Cloud, Cloud Shell e SDK. O capítulo conclui com uma discussão sobre diretrizes e boas práticas para gerenciar um cluster Kubernetes.

Capítulo 9: Computação com Cloud Run e App Engine

O Cloud Run e o App Engine fazem parte das ofertas serverless do Google Cloud.

Este capítulo introduz o Cloud Run, um serviço para executar contêineres na nuvem. Você aprenderá sobre a diferença entre Cloud Run Services e Cloud Run Jobs. O

Cloud Run provavelmente substituirá o App Engine como a escolha preferida para executar contêineres em um serviço serverless, mas o App Engine ainda está em uso e será abordado neste livro. Você aprenderá sobre componentes do App Engine, como aplicações, serviços, versões e instâncias. O capítulo também abrange como definir arquivos de configuração e especificar dependências de uma aplicação. Neste capítulo, você aprenderá como visualizar recursos do App Engine usando o console do Google Cloud, Cloud Shell e SDK. O capítulo também descreve como distribuir carga ajustando o tráfego com parâmetros de divisão. Você também aprenderá sobre autoescala no App Engine.

Capítulo 10: Computação com Cloud Functions

Cloud Functions é para cálculos sem servidor e orientados a eventos. Este capítulo introduz o Cloud Functions e mostra como usá-lo para receber eventos, evocar serviços e retornar resultados. Em seguida, você verá casos de uso para o Cloud Functions, como integração com APIs de terceiros e processamento orientado a eventos. Você aprenderá sobre o serviço Pub/Sub do Google para processamento baseado em publicação e subscrição e como usar o Cloud Functions com o Pub/Sub. Cloud Functions são bem adequados para responder a eventos no Cloud Storage. O capítulo descreve eventos do Cloud Storage e como usar o Cloud Functions para receber e responder a esses eventos. Você aprenderá a usar o Cloud Operations para monitorar e registrar detalhes das execuções do Cloud Function. Finalmente, o capítulo conclui com uma discussão sobre diretrizes para usar e gerenciar o Cloud Functions.

Capítulo 11: Planejando Armazenamento na Nuvem

Tendo descrito várias opções de computação no Google Cloud, é hora de voltar sua atenção para o armazenamento. Este capítulo descreve características dos sistemas de armazenamento, como seu tempo de acesso, persistência e modelo de dados. Neste capítulo, você aprenderá sobre as diferenças entre caches, armazenamento persistente e armazenamento arquivístico. Você aprenderá sobre as compensações de custo-benefício de usar armazenamento persistente regional e multirregional e usando armazenamento nearline versus Coldline e arquivístico. O capítulo inclui detalhes sobre as várias opções de armazenamento do Google Cloud, incluindo Cloud Storage para armazenamento de blob; Cloud SQL e Spanner para dados relacionais; Firestore e Bigtable, para armazenamento NoSQL; BigQuery para dados analíticos; e Cloud Firebase para dados de aplicativos móveis. O capítulo inclui orientações detalhadas sobre como escolher um armazenamento de dados com base em requisitos de consistência, disponibilidade, suporte a transações, custo, latência e suporte para vários padrões de leitura/escrita.

Capítulo 12: Implantando Armazenamento no Google Cloud Platform

Neste capítulo, você aprenderá como criar bancos de dados, adicionar dados, listar registros e excluir dados de cada um dos sistemas de armazenamento do Google Cloud. O capítulo começa introduzindo o Cloud SQL, um serviço de banco de dados gerenciado que oferece instâncias gerenciadas de SQL Server, MySQL e PostgreSQL. Você também aprenderá como criar bancos de dados no Cloud Firestore, BigQuery, Bigtable e Spanner. Em seguida, você voltará sua atenção para o Cloud Pub/Sub para armazenar dados em filas de mensagens, seguido por uma discussão sobre o Cloud Dataproc, um serviço

gerenciado de cluster Hadoop e Spark, para processamento de grandes conjuntos de dados. Na próxima seção, você aprenderá sobre o Cloud Storage para objetos. O capítulo conclui com orientações sobre como escolher um armazenamento de dados para um conjunto específico de requisitos.

Capítulo 13: Carregando Dados no Armazenamento

Existem várias maneiras de inserir dados no Google Cloud. Este capítulo descreve como usar o SDK de linha de comando para carregar dados no Cloud SQL, Cloud Storage, Firestore, BigQuery, Bigtable e Dataproc. Também descreve a importação e exportação em massa desses mesmos serviços. Em seguida, você aprenderá sobre dois padrões comuns de carregamento de dados: movendo dados do Cloud Storage e transmitindo dados para o Cloud Pub/Sub.

Capítulo 14: Redes na Nuvem: Nuvens Privadas Virtuais e Redes Privadas Virtuais

Neste capítulo, você voltará sua atenção para a rede com uma introdução aos conceitos básicos de rede, incluindo:

- Endereços IP
- Blocos CIDR
- Redes e sub-redes
- Nuvens privadas virtuais (VPCs)
- Roteamento e regras
- Redes privadas virtuais (VPNs)
- Cloud DNS
- Cloud Routers
- Cloud Interconnect
- Emparelhamento externo

Após ser apresentado aos conceitos-chave de rede, você aprenderá como criar uma VPC. Especificamente, isso inclui definir uma VPC, especificar regras de firewall, criar uma VPN e trabalhar com平衡adores de carga. Você aprenderá sobre diferentes tipos de平衡adores de carga e quando usá-los.

Capítulo 15: Redes na Nuvem: DNS, Balanceamento de Carga, Acesso Privado do Google e Endereçamento IP

Neste capítulo, você aprenderá sobre tarefas comuns de gerenciamento de rede, como definir sub-redes, adicionar subnets a uma VPC, gerenciar blocos CIDR e reservar endereços IP. Você aprenderá como realizar cada uma dessas tarefas usando o Cloud Console, Cloud Shell e Cloud SDK.

Capítulo 16: Implantando Aplicações com Cloud Marketplace e Cloud Foundation Toolkit

O Google Cloud Marketplace é o mercado do Google Cloud de pilhas e serviços pré-configurados. Este capítulo introduz o Cloud Marketplace e descreve algumas aplicações e serviços atualmente disponíveis. Você aprenderá como navegar pelo Cloud Marketplace, implantar aplicações a partir do Cloud Marketplace e encerrar aplicações do Cloud Marketplace. O capítulo também discute modelos de Deployment Manager que automatizam a implantação de uma aplicação e lançam um modelo do Deployment Manager para provisionar recursos do Google Cloud e configurar uma aplicação automaticamente.

Capítulo 17: Configurando Acesso e Segurança

Este capítulo introduz a gestão de identidade. Em particular, você aprenderá sobre identidades, papéis e como atribuir e remover papéis de identidade. Este capítulo também introduz contas de serviço e como criá-las, atribuí-las a VMs e trabalhar com elas em projetos. Você também aprenderá como visualizar registros de auditoria para projetos e serviços. O capítulo conclui com diretrizes para configurar a segurança do controle de acesso.

Capítulo 18: Monitoramento, Registro e Estimativa de Custos

No capítulo final, discutiremos alertas de Cloud Operations, registro, rastreamento distribuído e depuração de aplicativos. Cada um dos serviços correspondentes do Google Cloud é projetado para habilitar serviços mais eficientes, funcionais e confiáveis. O capítulo conclui com uma revisão do Calculador de Preços, que é útil para estimar o custo dos recursos no Google Cloud.

Ambiente de Aprendizado Interativo Online e Banco de Testes

Como todos os exames, a certificação de Engenheiro Associado da Nuvem do Google Cloud é atualizada periodicamente e pode eventualmente ser aposentada ou substituída. Em algum momento após o Google Cloud não oferecer mais este exame, as edições antigas de nossos livros e ferramentas online serão aposentadas. Se você comprou este livro após o exame ter sido aposentado, ou está tentando se registrar no ambiente de aprendizado online da Sybex após o exame ter sido aposentado, saiba que não garantimos que as ferramentas online da Sybex deste exame estarão disponíveis uma vez que o exame não estiver mais disponível.

Estudar o material no Guia de Estudo Certificado pelo Google Cloud Engenheiro Associado da Nuvem, Segunda Edição é uma parte importante da preparação para o exame de certificação de Engenheiro Associado da Nuvem, mas fornecemos ferramentas adicionais para ajudá-lo a se preparar. O Banco de Testes online ajudará você a entender os tipos de perguntas que aparecerão no exame de certificação. Os testes de amostra no Banco de Testes incluem todas as perguntas de cada capítulo, bem como as perguntas do teste de avaliação. Além disso, há dois exames práticos com 50 perguntas cada. Você pode usar esses testes para avaliar sua compreensão e identificar áreas onde pode precisar de estudo adicional.

Os flashcards no Banco de Testes desafiarão os limites do que você deve saber para o exame de certificação. São fornecidas 100 perguntas em formato digital. Cada flashcard tem uma pergunta e uma resposta correta.

O glossário online é uma lista pesquisável de termos-chave introduzidos neste guia de exame que você deve conhecer para o exame de certificação de Engenheiro Associado da Nuvem.

Para começar a usar esses para estudar para o exame de Engenheiro Associado da Nuvem Certificado pelo Google, vá para www.wiley.com/go/sybextestprep e registre seu livro para receber seu PIN único. Uma vez que você tenha o PIN, retorne a www.wiley.com/go/sybextestprep, encontre seu livro e clique em Registrar ou Login, e siga o link para registrar uma nova conta ou adicionar este livro a uma conta existente.

As políticas do exame podem mudar de tempos em tempos. Recomendamos fortemente que você verifique <https://cloud.google.com/certification> para as informações mais atualizadas quando começar sua preparação, quando se registrar e novamente alguns dias antes da data do seu exame agendado.

Objetivos do Exame

A certificação de Engenheiro de Nuvem Associado é projetada para pessoas que criam, implantam e gerenciam aplicações e infraestrutura empresarial no Google Cloud. Um Engenheiro de Nuvem Associado se sente confortável trabalhando com o Cloud Console, Cloud Shell e Cloud SDK. Esses indivíduos também entendem os produtos oferecidos como parte do Google Cloud e seus casos de uso apropriados.

O exame testará seu conhecimento do seguinte:

- Planejamento de uma solução na nuvem usando um ou mais serviços do Google Cloud
- Criação de um ambiente na nuvem para uma organização
- Implantação de aplicações e infraestrutura
- Uso de monitoramento e registro para garantir a disponibilidade das soluções na nuvem
- Configuração de gerenciamento de identidade, controles de acesso e outras medidas de segurança

Mapa de Objetivos

A seguir estão os objetivos específicos definidos pelo Google em <https://cloud.google.com/certification/guides/cloud-engineer>.

Seção 1: Configuração de um ambiente de solução na nuvem

1.1 Configuração de projetos e contas na nuvem. As atividades incluem:

Criação de uma hierarquia de recursos

Aplicação de políticas organizacionais à hierarquia de recursos

Concessão de papéis IAM a membros dentro de um projeto

Gerenciamento de usuários e grupos no Cloud Identity (manualmente e de forma automatizada)

Habilitação de APIs dentro dos projetos

Provisionamento e configuração de produtos na suíte de operações do Google Cloud

1.2 Gerenciamento da configuração de cobrança. As atividades incluem:

Criação de uma ou mais contas de cobrança

Vinculação de projetos a uma conta de cobrança

Estabelecimento de orçamentos e alertas de cobrança

Configuração de exportações de cobrança

1.3 Instalando e configurando a interface de linha de comando (CLI), especificamente o Cloud SDK (por exemplo, definindo o projeto padrão)

Seção 2: Planejando e configurando uma solução em nuvem

2.1 Planejando e estimando o uso de produtos do Google Cloud usando o Calculadora de Preços

2.2 Planejando e configurando recursos de computação. Considerações incluem:

■■ Selezionando opções de computação apropriadas para uma determinada carga de trabalho (por exemplo, Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)

■■ Usando VMs preemptivas e tipos de máquinas personalizadas conforme apropriado

2.3 Planejando e configurando opções de armazenamento de dados. Considerações incluem:

■■ Escolha do produto (por exemplo, Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)

■■ Escolhendo opções de armazenamento (por exemplo, disco persistente zonal, disco persistente balanceado regional, Standard, Nearline, Coldline, Archive)

2.4 Planejando e configurando recursos de rede. Tarefas incluem:

■■ Diferenciando opções de balanceamento de carga

■■ Identificando locais de recursos em uma rede para disponibilidade

■■ Configurando o Cloud DNS

Seção 3: Implementando e implantando uma solução em nuvem

3.1 Implementando e implantando recursos do Compute Engine. Tarefas incluem:

■■ Lançando uma instância de computação usando o Cloud Console e Cloud SDK (gcloud) (por exemplo, atribuir discos, política de disponibilidade, chaves SSH)

■■ Criando um grupo de instâncias gerenciadas escalonado automaticamente usando um modelo de instância

■■ Gerando/carregando uma chave SSH personalizada para instâncias

■■ Instalando e configurando o agente de Monitoramento e Logging do Cloud

■■ Avaliando quotas de computação e solicitando aumentos

3.2 Implementando e implantando recursos do Kubernetes Engine. Tarefas incluem:

■■ Instalando e configurando a interface de linha de comando (CLI) para Kubernetes (kubectl)

■■ Implantando um cluster do Google Kubernetes Engine com diferentes configurações, incluindo AutoPilot, clusters regionais, clusters privados, etc.

■■ Implantando uma aplicação conteinerizada no Google Kubernetes Engine

■■ Configurando o monitoramento e logging do Kubernetes Engine

3.3 Implantando e implementando recursos do Cloud Run e Cloud Functions. As tarefas incluem, quando aplicável:

- Implantação de uma aplicação e atualização da configuração de escalabilidade, versões e divisão de tráfego

- Implantação de uma aplicação que recebe eventos do Google Cloud (por exemplo, eventos Pub/Sub, eventos de notificação de mudança de objeto do Cloud Storage)

3.4 Implantando e implementando soluções de dados. As tarefas incluem:

- Inicialização de sistemas de dados com produtos (por exemplo, Cloud SQL, Firestore, BigQuery, Cloud Spanner, Cloud Pub/Sub, Cloud Bigtable, Dataproc, Dataflow, Cloud Storage)
- Carregamento de dados (por exemplo, upload via linha de comando, transferência via API, importação/exportação, carregar dados do Cloud Storage, transmitir dados para Pub/Sub)

3.5 Implantando e implementando recursos de rede. As tarefas incluem:

- Criação de um VPC com sub-redes (por exemplo, VPC em modo personalizado, VPC compartilhado)
- Lançamento de uma instância do Compute Engine com configuração de rede personalizada (por exemplo, endereço IP somente interno, acesso privado do Google, endereço IP externo e privado estático, tags de rede)
- Criação de regras de firewall de entrada e saída para um VPC (por exemplo, sub-redes IP, tags de rede, contas de serviço)
- Criação de uma VPN entre um VPC do Google e uma rede externa usando o Cloud VPN
- Criação de um balanceador de carga para distribuir o tráfego de rede de uma aplicação (por exemplo, balanceador de carga global HTTP(S), balanceador de carga Global SSL Proxy, balanceador de carga Global TCP Proxy, balanceador de carga de rede regional, balanceador de carga interno regional)

3.6 Implantando uma solução usando o Cloud Marketplace. As tarefas incluem:

- Navegação pelo catálogo do Cloud Marketplace e visualização dos detalhes da solução
- Implantação de uma solução do Cloud Marketplace

3.7 Implementando recursos via infraestrutura como código. As tarefas incluem:

- Construção de infraestrutura via templates do Cloud Foundation Toolkit e implementação de melhores práticas
- Instalação e configuração do Config Connector no Google Kubernetes Engine para criar, atualizar, deletar e proteger recursos

Seção 4: Garantindo a operação bem-sucedida de uma solução na nuvem 4.1 Gerenciando recursos do Compute Engine. As tarefas incluem:

- Gerenciamento de uma única instância de VM (por exemplo, iniciar, parar, editar configuração ou deletar uma instância)
- Conexão remota à instância

■■ Anexando uma GPU a uma nova instância e instalando as dependências necessárias

- Visualizando o inventário atual de VMs em execução (IDs de instâncias, detalhes)
- Trabalhando com snapshots (por exemplo, criar um snapshot de uma VM, visualizar snapshots, excluir um snapshot)
- Trabalhando com imagens (por exemplo, criar uma imagem a partir de uma VM ou snapshot, visualizar imagens, excluir uma imagem)
- Trabalhando com grupos de instâncias (por exemplo, definir parâmetros de autoescala, atribuir modelo de instância, criar um modelo de instância, remover um grupo de instâncias)
- Trabalhando com interfaces de gerenciamento (por exemplo, console do Google Cloud, Cloud Shell, Cloud SDK)

4.2 Gerenciando recursos do Kubernetes Engine. Tarefas incluem:

- Visualizando o inventário atual de clusters em execução (nós, pods, serviços)
- Navegando por imagens Docker e visualizando seus detalhes no Artifact Registry
- Trabalhando com grupos de nós (por exemplo, adicionar, editar ou remover um grupo de nós)
- Trabalhando com pods (por exemplo, adicionar, editar ou remover pods)
- Trabalhando com serviços (por exemplo, adicionar, editar ou remover um serviço)
- Trabalhando com aplicações stateful (por exemplo, volumes persistentes, conjuntos stateful)
- Gerenciando configurações de autoescala horizontal e vertical
- Trabalhando com interfaces de gerenciamento (por exemplo, console do Google Cloud, Cloud Shell, Cloud SDK, kubectl)

4.3 Gerenciando recursos do Cloud Run. Tarefas incluem:

- Ajustando parâmetros de divisão de tráfego da aplicação
- Definindo parâmetros de escala para instâncias de autoescala
- Determinando se deve executar o Cloud Run (totalmente gerenciado) ou o Cloud Run para Anthos

4.4 Gerenciando soluções de armazenamento e banco de dados. Tarefas incluem:

- Gerenciando e protegendo objetos dentro e entre buckets do Cloud Storage
- Definindo políticas de ciclo de vida de objetos para buckets do Cloud Storage
- Executando consultas para recuperar dados de instâncias de dados (por exemplo, Cloud SQL, BigQuery, Cloud Spanner, Datastore, Cloud Bigtable)
- Estimando custos de recursos de armazenamento de dados

■■ Fazendo backup e restaurando instâncias de banco de dados (por exemplo, Cloud SQL, Datastore)

■■ Revisando o status de trabalhos no Dataproc, Dataflow ou BigQuery

4.5 Gerenciando recursos de rede. As tarefas incluem:

- Adicionando uma sub-rede a um VPC existente
- Expandindo uma sub-rede para ter mais endereços IP
- Reservando endereços IP externos ou internos estáticos
- Trabalhando com CloudDNS, CloudNAT, Balanceadores de Carga e regras de firewall

4.6 Monitoramento e registro. As tarefas incluem:

- Criando alertas do Cloud Monitoring baseados em métricas de recursos
- Criando e ingerindo métricas personalizadas do Cloud Monitoring (por exemplo, de aplicações ou logs)
- Configurando sinks de logs para exportar logs para sistemas externos (por exemplo, locais ou BigQuery)
- Configurando roteadores de logs
- Visualizando e filtrando logs no Cloud Logging
- Visualizando detalhes específicos de mensagens de log no Cloud Logging
- Usando diagnósticos na nuvem para pesquisar um problema de aplicação (por exemplo, visualizando dados do Cloud Trace, usando o Cloud Debug para visualizar um ponto específico no tempo de uma aplicação)
- Visualizando o status do Google Cloud

Seção 5: Configurando acesso e segurança 5.1 Gerenciando o Gerenciamento de Identidade e Acesso (IAM). As tarefas incluem:

- Visualizando políticas IAM
- Criando políticas IAM
- Gerenciando os vários tipos de papel e definindo papéis IAM personalizados (por exemplo, primitivos, predefinidos e personalizados)

5.2 Gerenciando contas de serviço. As tarefas incluem:

- Criando contas de serviço
- Usando contas de serviço em políticas IAM com permissões mínimas
- Atribuindo contas de serviço a recursos
- Gerenciando o IAM de uma conta de serviço

- Gerenciando a personificação de conta de serviço
- Criando e gerenciando credenciais de conta de serviço de curta duração

5.3 Visualizando registros de auditoria

Teste de Avaliação

1. Modelos de instância são usados para criar um grupo de VMs idênticas. Os modelos de instância incluem:
 - A. Tipo de máquina, imagem de disco de inicialização ou imagem de contêiner, zona e etiquetas
 - B. Definições de bucket do Cloud Storage
 - C. Uma descrição de balanceador de carga
 - D. Arquivo de configuração do App Engine
2. O comando de linha de comando para criar um bucket do Cloud Storage é:
 - A. gcloud mb
 - B. gsutil mb
 - C. gcloud mkbucket
 - D. gsutil mkbucket
3. Sua empresa tem uma política de gestão de objetos que exige que objetos armazenados no Cloud Storage sejam migrados do armazenamento padrão para o armazenamento nearline 90 dias após a criação do objeto. A maneira mais eficiente de fazer isso é:
 - A. Criar uma Cloud Function para copiar objetos do armazenamento regional para o armazenamento nearline.
 - B. Definir a propriedade MigrateObjectAfter no objeto armazenado para 90 dias.
 - C. Copiar o objeto para o armazenamento persistente anexado a uma VM e depois copiar o objeto para um bucket criado no armazenamento nearline.
 - D. Criar uma política de configuração de gestão de ciclo de vida especificando uma idade de 90 dias e SetStorageClass como nearline.
4. Um cliente da área de educação mantém um site onde os usuários podem fazer upload de vídeos, e seu cliente precisa garantir a redundância dos arquivos; portanto, você criou dois buckets para o Cloud Storage. Qual comando você usa para sincronizar o conteúdo dos dois buckets?
 - A. gsutil rsync
 - B. gcloud cp sync
 - C. gcloud rsync
 - D. gsutil cp sync
5. Os recursos de VPC são quais dos seguintes?

A. Regional

B. Zonal

C. Global

D. Sub-rede

6. Um componente remoto na sua rede falhou, o que resulta em um erro de rede transitório. Quando você executa um comando gsutil, ele falha devido a um erro transitório. Por padrão, o comando irá:

A. Terminar e registrar uma mensagem no Cloud Monitoring

B. Tentar novamente usando uma estratégia de recuo binário exponencial truncado

C. Perguntar ao usuário se deseja tentar novamente ou sair

D. Terminar e registrar uma mensagem no Cloud Shell

7. Todos os seguintes são componentes de regras de firewall, exceto qual?

A. Direção do tráfego

B. Ação na correspondência

C. Tempo de vida (TTL)

D. Protocolo

8. Adicionar máquinas virtuais a um grupo de instâncias pode ser disparado em uma política de autoescala por todos os seguintes, exceto qual?

A. Utilização da CPU

B. Métricas do Cloud Monitoring

C. Violação de política do IAM

D. Capacidade de atendimento do balanceamento de carga

9. O departamento financeiro da sua empresa está desenvolvendo uma nova aplicação de gerenciamento de contas que requer transações e a capacidade de realizar operações de banco de dados relacional usando SQL totalmente compatível. As opções de armazenamento de dados no Google Cloud incluem:

A. Spanner e Cloud SQL

B. Firestore e Bigtable

C. Spanner e Cloud Storage

D. Firestore e Cloud SQL

10. O departamento de marketing da sua empresa deseja implantar uma aplicação web mas não quer ter que gerenciar servidores ou clusters. Uma boa opção para eles é:

A. Compute Engine

- B. Kubernetes Engine
 - C. Cloud Run
 - D. Cloud Functions
11. Sua empresa está construindo um data warehouse empresarial e deseja capacidades de consulta SQL sobre petabytes de dados, mas não quer gerenciar servidores ou clusters. Uma boa opção para eles é:
- A. Cloud Storage
 - B. BigQuery
 - C. Bigtable
 - D. Firestore
12. Você foi contratado como consultor para uma startup na área de Internet das Coisas (IoT). A startup transmitirá grandes volumes de dados para o Google Cloud. Os dados precisam ser filtrados, transformados e analisados antes de serem armazenados no Google Cloud Firestore. Uma boa opção para o componente de processamento de fluxo é:
- A. Dataproc
 - B. Cloud Dataflow
 - C. Cloud Endpoints
 - D. Cloud Interconnect
13. Máquinas virtuais preemptíveis podem ser desligadas a qualquer momento, mas sempre serão desligadas após funcionarem:
- A. 6 horas
 - B. 12 horas
 - C. 24 horas
 - D. 48 horas
14. Você recebeu a tarefa de projetar uma hierarquia organizacional para gerenciar departamentos e seus recursos na nuvem. Quais componentes de organização estão disponíveis no Google Cloud?
- A. Organizaçāo, pastas, projetos
 - B. Buckets, diretórios, subdiretórios
 - C. Organizações, buckets, projetos
 - D. Pastas, buckets, projetos

15. Durante um incidente que causou a falha de uma aplicação, você suspeita que algum recurso pode não ter as funções apropriadas concedidas. O comando para listar funções concedidas a um recurso é:

- A. gutil iam list-grantable-roles
- B. gcloud iam list-grantable-roles
- C. gcloud list-grantable-roles
- D. gcloud resources grantable-roles

16. A disponibilidade de plataformas de CPU pode variar entre zonas. Para obter uma lista de todos os tipos de CPU disponíveis em uma zona específica, você deve usar:

- A. gcloud compute zones describe
- B. gcloud iam zones describe
- C. gutil zones describe
- D. gcloud compute regions list

17. Para criar uma função personalizada, um usuário deve possuir qual função?

- A. iam.create
- B. compute.roles.create
- C. iam.roles.create
- D. Compute.roles.add

18. Foi solicitado que você crie uma rede com 1.000 endereços IP. Com o objetivo de minimizar endereços IP não utilizados, qual sufixo CIDR você usaria para criar uma rede com pelo menos 1.000 endereços, mas não mais do que o necessário?

- A. /20
- B. /22
- C. /28
- D. /32

19. Um time de cientistas de dados pediu sua ajuda para configurar um cluster Apache Spark. Você sugere que eles usem um serviço gerenciado do Google Cloud em vez de gerenciar um cluster por conta própria no Compute Engine. O serviço que eles usariam é:

- A. Dataproc
- B. Cloud Dataflow
- C. Cloud Hadoop
- D. BigQuery

20. Você criou uma aplicação web que permite aos usuários fazer upload de arquivos para o Cloud Storage. Quando os arquivos são carregados, você quer verificar o tamanho do arquivo e atualizar o total de armazenamento usado na conta do usuário. Uma opção sem servidor para realizar essa ação no carregamento é:
- A. Cloud Dataflow
 - B. Dataproc
 - C. Cloud Storage
 - D. Cloud Functions
21. Sua empresa acabou de começar a usar o Google Cloud, e os executivos querem ter uma conexão dedicada do seu data center para o Google Cloud para permitir grandes transferências de dados. Qual serviço de rede você recomendaria?
- A. Google Cloud Carrier Internet Peering
 - B. Google Cloud Dedicated Interconnect
 - C. Google Cloud Internet Peering
 - D. Google Cloud DNS
22. Você quer que o Google Cloud gerencie chaves criptográficas, então decidiu usar o Cloud Key Management Services. Antes de você começar a criar chaves criptográficas, você deve:
- A. Ativar a API do Google Cloud Key Management Service (KMS) e configurar a cobrança.
 - B. Ativar a API do Google Cloud KMS e criar pastas.
 - C. Criar pastas e configurar a cobrança.
 - D. Dar a todos os usuários papéis concedíveis para criar chaves.
23. No Kubernetes Engine, um pool de nós é:
- A. Um subconjunto de nós entre clusters
 - B. Um conjunto de VMs gerenciadas fora do Kubernetes Engine
 - C. Um conjunto de VMs preemptíveis
 - D. Um subconjunto de instâncias de nós dentro de um cluster que têm a mesma configuração
24. O serviço do Google Cloud para armazenar e gerenciar contêineres Docker é:
- A. Cloud DevOps Repository
 - B. Cloud Build
 - C. Container Registry

D. Docker Repository

25. O código para Cloud Functions pode ser escrito em várias linguagens, incluindo:

- A. Apenas Node.js e Python
- B. Node.js, Python e Go
- C. Python e Go
- D. Python e C

Respostas do Teste de Avaliação

1. A. Tipo de máquina, imagem de disco de inicialização ou imagem de contêiner, zona e rótulos são todos parâmetros de configuração ou atributos de uma VM e, portanto, seriam incluídos em uma configuração de grupo de instâncias que cria essas VMs.
2. B. gsutil é a linha de comando para acessar e manipular o Cloud Storage a partir da linha de comando. mb é o comando específico para criar, ou fazer, um bucket.
3. D. A política de configuração de ciclo de vida permite especificar critérios para migrar dados para outros sistemas de armazenamento sem ter que se preocupar em executar trabalhos para executar as etapas necessárias. As outras opções são ineficientes ou não existem.
4. A. gsutil é a ferramenta de linha de comando para trabalhar com o Cloud Storage. rsync é o comando específico no gsutil para sincronizar buckets.
5. C. O Google opera uma rede global, e as VPCs são recursos que podem abranger essa rede global.
6. B. gcloud por padrão tentará novamente uma operação de rede falha e esperará muito tempo antes de cada tentativa. O tempo de espera é calculado usando uma estratégia de backoff exponencial binário truncado.
7. C. Regras de firewall não têm parâmetros TTL. Direção do tráfego, ação em correspondência e protocolo são todos componentes das regras de firewall.
8. C. Violações de política do IAM não acionam mudanças no tamanho dos clusters. Todas as outras opções podem ser usadas para acionar uma mudança no tamanho do cluster.
9. A. Apenas os bancos de dados Spanner e Cloud SQL suportam transações e têm uma interface SQL. O Firestore tem transações, mas não suporta SQL totalmente compatível; possui uma linguagem de consulta semelhante ao SQL. O Cloud Storage não suporta transações ou SQL.
10. C. Cloud Run é um serviço sem servidor para executar contêineres e permite que desenvolvedores implantem aplicações completas sem ter que gerenciar servidores ou clusters. Compute Engine e Kubernetes Engine requerem gerenciamento de servidores. O Cloud Functions é adequado para funções de curta duração em Node.js ou Python, mas não para aplicações completas.
11. B. O BigQuery é projetado para análises em escala de petabytes e fornece uma interface SQL.
12. B. O Cloud Dataflow permite o processamento de fluxo e em lote de dados e é bem adequado para esse tipo de trabalho ETL. O Dataproc é um serviço gerenciado Hadoop e Spark usado para análises de big data. O Cloud Endpoints é um serviço de API, e o Cloud Interconnect é um serviço de rede.

13. C. Se uma máquina preemptível não for desligada dentro de 24 horas, o Google interromperá a instância.
14. A. Organizações, pastas e projetos são os componentes usados para gerenciar uma hierarquia organizacional. Buckets, diretórios e subdiretórios são usados para organizar o armazenamento.
15. B. gcloud é a ferramenta de linha de comando para trabalhar com o IAM, e list-grantable-roles é o comando correto.
16. A. gcloud é a ferramenta de linha de comando para manipular recursos de computação, e zones describe é o comando correto.
17. C. iam.roles.create está correto; os outros papéis não existem.
18. B. O sufixo /22 produz 1.022 endereços IP utilizáveis.
19. A. Dataproc é o serviço gerenciado para Spark. Cloud Dataflow é para processamento de dados em stream e lote, BigQuery é para análise, e Cloud Hadoop não é um serviço do Google Cloud.
20. D. Cloud Functions responde a eventos no Cloud Storage, tornando-o uma boa escolha para tomar uma ação após um arquivo ser carregado.
21. B. Google Cloud Dedicated Interconnect é a única opção para uma conexão dedicada entre o data center de um cliente e um data center do Google.
22. A. Ativar a API do Google Cloud KMS e configurar a cobrança são etapas comuns ao usar serviços do Google Cloud.
23. D. Um pool de nós é um subconjunto de instâncias de nós dentro de um cluster que têm todas a mesma configuração.
24. C. O serviço do Google Cloud para armazenar e gerenciar contêineres Docker é o Artifact Registry. Cloud Build é para criar imagens. Cloud Source Repositories são repositórios Git privados hospedados no Google Cloud. Docker Repository não é um serviço do Google Cloud.
25. B. Node.js, Python e Go são três das linguagens suportadas pelo Cloud Functions.

Capítulo 1

Visão Geral do Google Cloud

ESTE CAPÍTULO COBRE O SEGUINTE OBJETIVO DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 1.0 Configuração de projetos e contas na nuvem

O Google Cloud é um serviço de nuvem pública que oferece algumas das mesmas tecnologias usadas pelo Google para entregar seus próprios produtos. Este capítulo descreve os componentes mais importantes do Google Cloud e discute como ele difere da computação baseada em data centers locais (on-premises).

Tipos de Serviços na Nuvem

Provedores de nuvem pública como Google, Amazon e Microsoft oferecem uma gama de serviços para a implantação de computação, armazenamento, rede e outras infraestruturas para executar uma ampla variedade de serviços empresariais e aplicações. Alguns usuários da nuvem são novas empresas que começam na nuvem. Eles nunca possuíram sua própria infraestrutura de hardware. Outros clientes da nuvem são empresas com múltiplos data centers que usam nuvens públicas para complementar seus data centers. Esses diferentes tipos de usuários têm requisitos diferentes.

Uma empresa que começa na nuvem pode escolher serviços que melhor se adequem às suas necessidades de aplicação e arquitetura sem ter que considerar a infraestrutura existente. Por exemplo, uma startup poderia usar o Cloud Identity do Google Cloud, incluindo serviços de Gerenciamento de Identidade e Acesso, para todas as necessidades de autenticação e autorização. Uma empresa que já investiu em uma solução Microsoft Active Directory para gerenciamento de identidade pode querer aproveitar esse sistema em vez de trabalhar exclusivamente com o sistema de gerenciamento de identidade da nuvem. Isso pode levar a trabalho adicional para integrar os dois sistemas e mantê-los sincronizados.

Outra área de preocupação para empresas com sua própria infraestrutura é estabelecer e manter uma rede segura entre seus recursos locais e seus recursos na nuvem pública. Se houver um tráfego de rede de alto volume entre os sistemas locais e a nuvem pública, a empresa pode precisar investir em rede dedicada entre seu data center e uma instalação do provedor de nuvem pública. Se o volume de tráfego não justificar o custo de uma conexão dedicada entre instalações, então a empresa pode usar uma rede privada virtual que opera sobre a Internet pública. Isso requer design de rede adicional e gerenciamento que uma empresa que está exclusivamente na nuvem não precisaria abordar.

Provedores de nuvem pública oferecem serviços que se enquadram em quatro categorias amplas:

- Recursos de computação
- Armazenamento
- Rede
- Serviços especializados como serviços de aprendizado de máquina

Clientes de nuvem tipicamente fazem uso de serviços em mais de uma dessas categorias.

Recursos Computacionais

Os recursos computacionais assumem várias formas nas nuvens públicas.

Máquinas Virtuais

Máquinas virtuais (VMs) são uma unidade básica de recursos computacionais e um bom ponto de partida para experimentar a nuvem. Após criar uma conta com um

provedor de nuvem e fornecer informações de cobrança, você pode criar um projeto, que é um agrupamento lógico de recursos do Google Cloud. Uma vez que você tenha um projeto, pode usar um portal ou ferramentas de linha de comando para criar VMs nesse projeto. O Google Cloud oferece uma variedade de VMs pré-configuradas com diferentes números de vCPUs e quantidades de memória. Você também pode criar uma configuração personalizada caso as ofertas pré-configuradas não atendam às suas necessidades.

Uma vez que você cria uma VM, pode fazer login nela e administrá-la como desejar. Você tem acesso total à VM, então pode configurar sistemas de arquivos, adicionar armazenamento persistente, atualizar o sistema operacional ou instalar pacotes adicionais. Você decide o que executar na VM, quem mais terá acesso a ela e quando desligar a VM. Uma VM que você gerencia é como ter um servidor em seu escritório ao qual você tem plenos direitos administrativos.

Você pode, claro, criar várias VMs executando diferentes sistemas operacionais e aplicações. O Google Cloud também fornece serviços, como平衡adores de carga, que fornecem um único ponto de acesso a um back-end distribuído. Isso é especialmente útil quando você precisa ter alta disponibilidade para sua aplicação. Se uma das VMs em um cluster falhar, a carga de trabalho pode ser direcionada para as outras VMs no cluster. Autoescaladores podem adicionar ou remover VMs do cluster baseado na carga de trabalho. Isso é chamado de autoescala. Isso ajuda tanto a controlar custos, por não executar mais VMs do que o necessário, quanto a garantir que capacidade computacional suficiente esteja disponível quando as cargas de trabalho aumentam.

Clusters Kubernetes Gerenciados

O Google Cloud oferece todas as ferramentas necessárias para criar e gerenciar clusters de servidores. Muitos usuários da nuvem prefeririam focar em suas aplicações e não nas tarefas necessárias para manter um cluster de servidores funcionando. Para esses usuários, os clusters gerenciados são uma boa opção. Os clusters gerenciados fazem uso de contêineres. Um contêiner é como uma VM leve que isola processos executados em um contêiner de processos executados em outro contêiner no mesmo servidor. Em um cluster gerenciado, você pode especificar o número de servidores que gostaria de executar e os contêineres que devem ser executados neles. Você também pode especificar parâmetros de autoescala para otimizar o número de contêineres em execução.

Em um cluster gerenciado, a saúde dos contêineres é monitorada para você. Se um contêiner falhar, o software de gerenciamento do cluster detectará isso e iniciará outro contêiner.

Contêineres são boas opções quando você precisa executar aplicações que dependem de vários microsserviços executando em seu ambiente. Os serviços são implantados por meio de contêineres, e o serviço de gerenciamento de cluster cuida do monitoramento, rede e algumas tarefas de gerenciamento de segurança.

Computação Sem Servidor (Serverless)

Tanto VMs quanto clusters Kubernetes gerenciados requerem algum nível de esforço para configurar e administrar recursos de computação. A computação sem servidor é uma abordagem que permite que desenvolvedores e administradores de

aplicações executem seu código em um ambiente computacional que não requer a configuração de VMs ou clusters Kubernetes.

O Google Cloud tem três opções de computação sem servidor: App Engine, Cloud Run e Cloud Functions. O App Engine é usado para aplicações e contêineres que rodam por períodos prolongados, como um backend de site, sistema de ponto de venda ou aplicação comercial personalizada. O Cloud Run também é usado para executar contêineres quando as funcionalidades completas do Kubernetes Engine não são necessárias. O Cloud Run é usado para contêineres quando você deseja um serviço totalmente gerenciado e autoescala rápida para suas aplicações sem estado. Cloud Functions é uma plataforma para executar código em resposta a um evento, como o upload de um arquivo ou a adição de uma mensagem a uma fila de mensagens. Esta opção sem servidor funciona bem quando você precisa responder a um evento executando um processo curto codificado em uma função ou chamando uma aplicação de longa duração que pode estar rodando em uma VM, cluster gerenciado ou App Engine.

Armazenamento

Nuvens públicas oferecem alguns tipos de serviços de armazenamento úteis para uma ampla gama de requisitos de aplicação. Esses tipos incluem o seguinte:

- Armazenamento de objetos
- Armazenamento de arquivos
- Armazenamento de blocos
- Caches

Usuários empresariais de serviços de nuvem frequentemente usam uma combinação desses serviços.

Armazenamento de Objetos

Armazenamento de objetos é um sistema que gerencia o uso de armazenamento em termos de objetos ou blobs. Geralmente, esses objetos são arquivos, mas é importante notar que os arquivos não são armazenados em um sistema de arquivos convencional. Os objetos são agrupados em buckets (recipientes). Cada objeto é endereçável individualmente, geralmente por uma URL.

O armazenamento de objetos não é limitado pelo tamanho de discos ou unidades de estado sólido (SSDs) conectados a um servidor. Objetos podem ser carregados sem preocupação com a quantidade de espaço disponível em um disco. Múltiplas cópias dos objetos são armazenadas para melhorar a disponibilidade e durabilidade. Em alguns casos, cópias dos objetos podem ser armazenadas em diferentes regiões para garantir a disponibilidade mesmo se uma região se tornar inacessível.

Outra vantagem do armazenamento de objetos é que ele é sem servidor. Não há necessidade de criar VMs e anexar armazenamento a elas. O armazenamento de objetos do Google Cloud, chamado Cloud Storage, é acessível a partir de servidores executando no Google Cloud, bem como de outros dispositivos com acesso à Internet.

Armazenamento de Arquivos

Os serviços de armazenamento de arquivos fornecem um sistema de armazenamento hierárquico para arquivos. O armazenamento de sistema de arquivos oferece sistemas de arquivos compartilhados na rede. O Google Cloud tem um serviço de armazenamento de arquivos chamado Cloud Filestore, que é baseado no sistema de armazenamento Network File System (NFS).

O armazenamento de arquivos é adequado para aplicações que requerem acesso a arquivos semelhante ao de sistemas operacionais. O sistema de armazenamento de arquivos desacopla o sistema de arquivos de VMs específicas. O sistema de arquivos, seus diretórios e seus arquivos existem independentemente de VMs ou aplicações que possam acessar esses arquivos.

Armazenamento de Blocos

O armazenamento de blocos usa uma estrutura de dados de tamanho fixo chamada bloco para organizar dados. O armazenamento de blocos é comumente usado em discos efêmeros e persistentes anexados a VMs. Com um sistema de armazenamento de blocos, você pode instalar sistemas de arquivos em cima do armazenamento de blocos, ou pode executar aplicações que acessam os blocos diretamente. Alguns bancos de dados relacionais podem ser projetados para acessar blocos diretamente, em vez de trabalhar por meio de sistemas de arquivos.

Em sistemas de arquivos Linux, 4 KB é um tamanho comum de bloco. Bancos de dados relacionais frequentemente escrevem diretamente em blocos, mas muitas vezes usam tamanhos maiores, como 8 KB ou mais.

O armazenamento de blocos está disponível em discos que são anexados a VMs no Google Cloud. O armazenamento de blocos pode ser tanto persistente quanto efêmero. Um disco persistente continua a existir e armazenar dados, mesmo se for desanexado de um servidor virtual ou o servidor virtual ao qual está anexado for desligado. Discos efêmeros existem e armazenam dados apenas enquanto uma VM está em execução. Discos efêmeros são excluídos quando a VM é desligada. Discos persistentes são usados quando você quer que os dados existam em um dispositivo de armazenamento de blocos independentemente de uma VM. Esses discos são boas opções quando você tem dados que quer disponíveis independentemente do ciclo de vida de uma VM e suportam acesso rápido a nível de sistema operacional e sistema de arquivos.

O armazenamento de objetos também mantém dados independentes do ciclo de vida de uma VM, mas não suporta acesso a nível de sistema operacional ou sistema de arquivos; você tem que usar protocolos de nível mais alto como HTTP para acessar objetos. Leva mais tempo para recuperar dados do armazenamento de objetos do que recuperá-los do armazenamento de blocos. Você pode precisar de uma combinação de armazenamento de objetos e armazenamento de blocos para atender às suas necessidades de aplicação. O armazenamento de objetos pode armazenar grandes volumes de dados que são copiados para o disco persistente quando necessário. Esta combinação oferece a vantagem de grandes volumes de armazenamento juntamente com acesso baseado em sistema operacional e sistema de arquivos quando necessário.

Caches

Caches são armazenamentos de dados em memória que mantêm um acesso rápido aos dados. O tempo necessário para recuperar os dados é chamado de latência. A latência de armazenamentos em memória é projetada para ser submilissegundos. Para lhe dar uma comparação, aqui estão algumas outras latências:

- Fazer uma referência à memória principal leva 100 nanosegundos, ou 0,1 microsegundo.
- Ler 4 KB aleatoriamente de um SSD leva 150 microsegundos.
- Ler 1 MB sequencialmente da memória leva 250 microsegundos.
- Ler 1 MB sequencialmente de um SSD leva 1.000 microsegundos, ou 1 milissegundo.
- Ler 1 MB sequencialmente de um disco leva 20.000 microsegundos, ou 20 milissegundos.

Aqui estão algumas conversões para referência:

- 1.000 nanosegundos igual a 1 microsegundo.
- 1.000 microsegundos igual a 1 milissegundo.
- 1.000 milissegundos igual a 1 segundo.

Esses e outros dados de tempo úteis estão disponíveis em “Latency Numbers Every Programmer Should Know” de Jonas Bonér em <https://gist.github.com/jboner/2841832>.

Vamos trabalhar através de um exemplo de leitura de 1 MB de dados. Se você tiver os dados armazenados em um cache em memória, você pode recuperar os dados em 250 microsegundos, ou 0,25 milissegundo. Se esses mesmos dados estiverem armazenados em um SSD, levará quatro vezes mais tempo para recuperá-los, ou seja, 1 milissegundo. Se você recuperar os mesmos dados de um disco rígido (HDD), pode esperar esperar 20 milissegundos, ou 80 vezes mais tempo do que a leitura de um cache em memória.

Caches são bastante úteis quando você precisa manter a latência de leitura ao mínimo em sua aplicação. Claro, quem não adora tempos de recuperação rápidos? Por que não armazenamos sempre nossos dados em caches? Há três razões:

- A memória é mais cara do que o armazenamento em SSD ou HDD. Não é prático, em muitos casos, ter tanto armazenamento em memória quanto armazenamento em bloco persistente em SSDs ou HDDs.
- Caches são voláteis; você perde os dados armazenados no cache quando a energia é cortada ou o sistema operacional é reiniciado. Você pode armazenar dados em um cache para acesso rápido, mas ele nunca deve ser usado como o único repositório de dados mantendo os dados. Alguma forma de armazenamento persistente deve ser usada para manter um "sistema da verdade", ou um repositório de dados que sempre tem a versão mais recente e mais precisa dos dados.

- Caches podem ficar desincronizados com o sistema da verdade. Isso pode acontecer se o sistema da verdade for atualizado, mas os novos dados não forem escritos no cache. Quando isso acontece, pode ser difícil para uma aplicação que depende do cache detectar o fato de que os dados no cache são inválidos. Se você decidir usar um cache, certifique-se de projetar uma estratégia de atualização de cache que atenda às suas exigências de consistência entre o cache e o sistema da verdade. Esse é um problema de design tão desafiador que se tornou imortalizado na conhecida observação de Phil Karlton: "Há apenas duas coisas difíceis em ciência da computação: invalidação de cache e nomear coisas." (Veja <https://martinfowler.com/bliki/TwoHardThings.html> para variações deste raro exemplo de humor em ciência da computação.)

Cenário Real

Melhorando o Tempo de Resposta de Consultas em Banco de Dados Os usuários esperam que as aplicações web sejam altamente responsivas. Se uma página leva mais de 2 a 3 segundos para carregar, a experiência do usuário pode ser afetada. É comum gerar o conteúdo de uma página usando os resultados de uma consulta ao banco de dados, como procurar informações de conta por ID do cliente. Quando uma consulta é feita ao banco de dados, o motor do banco de dados procura os dados, que geralmente estão em disco. Quanto mais usuários consultam o banco de dados, mais consultas ele tem que atender. Os bancos de dados mantêm uma fila para consultas que precisam ser respondidas, mas não podem ser processadas ainda porque o banco de dados está ocupado com outras consultas. Isso pode causar um tempo de resposta de latência maior, já que a aplicação web terá que esperar o banco de dados retornar os resultados da consulta.

Uma maneira de reduzir a latência é reduzir o tempo necessário para ler os dados. Em alguns casos, ajuda substituir discos rígidos por drives SSD mais rápidos. No entanto, se o volume de consultas é alto o suficiente que a fila de consultas é longa mesmo com SSDs, outra opção é usar um cache.

Quando os resultados das consultas são buscados, eles são armazenados no cache. Na próxima vez que essa informação for necessária, ela é buscada do cache em vez do banco de dados. Isso pode reduzir a latência porque os dados são buscados da memória, que é mais rápida do que o disco. Também reduz o número de consultas ao banco de dados, então consultas que não podem ser respondidas procurando dados no cache não terão que esperar tanto na fila de consultas antes de serem processadas.

Essa abordagem exigiria mudanças no código da aplicação para armazenar os resultados das consultas no cache e verificar o cache por dados antes de consultar o banco de dados.

Redes

Ao trabalhar na nuvem, você precisará lidar com a rede entre seus recursos na nuvem e, possivelmente, com seus sistemas locais.

Quando você tem várias VMs executando em seu ambiente de nuvem, provavelmente precisará gerenciar endereços IP em algum momento. Cada dispositivo ou serviço acessível por rede em seu ambiente precisará de um endereço IP. De fato,

dispositivos dentro do Google Cloud podem ter tanto endereços internos quanto externos. Endereços internos são acessíveis apenas por serviços em sua rede interna. Sua rede interna do Google Cloud é definida como uma nuvem privada virtual (VPC). Endereços externos são acessíveis da Internet.

Endereços IP externos podem ser estáticos ou efêmeros. Endereços estáticos são atribuídos a um dispositivo por períodos prolongados de tempo. Endereços IP externos efêmeros são anexados a VMs e liberados quando a VM é parada.

Além de especificar endereços IP, você frequentemente precisará definir regras de firewall para controlar o acesso a sub-redes e VMs em sua VPC. Por exemplo, você pode ter um servidor de banco de dados ao qual deseja restringir o acesso, de modo que apenas um servidor de aplicação possa consultar o banco de dados. Uma regra de firewall pode ser configurada para limitar o tráfego de entrada e saída para o endereço IP do servidor de aplicação ou balanceador de carga à frente do cluster de aplicação.

Você pode precisar compartilhar dados e acesso à rede entre um data center local e sua VPC. Você pode fazer isso usando um dos vários tipos de emparelhamento, que é o termo geral para a ligação de redes distintas. O Google Cloud oferece vários tipos de emparelhamento, incluindo VPNs, Interconexões, VPC compartilhada, emparelhamento de redes VPC e emparelhamento Direto ou por Operadora.

Serviços Especializados

A maioria dos provedores de nuvem pública oferece serviços especializados que podem ser usados como blocos de construção de aplicações ou como parte de um fluxo de trabalho para processamento de dados. Características comuns dos serviços especializados incluem:

- Eles são sem servidor; você não precisa configurar servidores ou clusters.
- Eles fornecem uma função específica, como traduzir textos ou analisar imagens.
- Eles fornecem uma interface de programação de aplicativos (API) para acessar a funcionalidade do serviço.
- Assim como outros serviços de nuvem, você é cobrado com base no seu uso do serviço.

Aqui estão alguns dos serviços especializados no Google Cloud:

- AutoML, um serviço de aprendizado de máquina
- Cloud Natural Language, um serviço para análise de texto
- Speech-to-Text para converter a linguagem falada em texto
- Recommendations AI para recomendações personalizadas de produtos

Serviços especializados encapsulam capacidades computacionais avançadas e as tornam acessíveis a desenvolvedores que não são especialistas em domínios, como processamento de linguagem natural e aprendizado de máquina. Espere ver mais serviços especializados sendo adicionados ao Google Cloud.

Cloud Computing vs. Data Center Computing

Embora possa parecer que executar VMs na nuvem não é muito diferente de executá-las no seu centro de dados, existem diferenças significativas entre operar ambientes de TI na nuvem e um centro de dados local ou colocado em colocation.

Alugar em Vez de Possuir Recursos

Os data centers corporativos estão cheios de servidores, arrays de disco e equipamentos de rede. Este equipamento é frequentemente possuído ou alugado por períodos prolongados pela empresa, um modelo que exige que as empresas gastem uma quantia significativa de dinheiro antecipadamente para comprar equipamentos ou se comprometam com um aluguel de longo prazo para o equipamento. Esta abordagem funciona bem quando uma organização pode prever com precisão o número de servidores e outros equipamentos de que precisará por um período prolongado e pode utilizar esse equipamento de forma consistente.

O modelo não funciona tão bem quando as empresas têm que planejar para uma capacidade de pico significativamente maior do que a carga de trabalho média. Por exemplo, um varejista pode ter uma carga média que requer um cluster de 20 servidores, mas durante a temporada de festas a carga de trabalho aumenta ao ponto de serem necessários 80 servidores. A empresa poderia comprar 80 servidores e deixar 60 ociosos durante a maior parte do ano para ter recursos para acomodar a capacidade de pico. Alternativamente, poderia comprar ou alugar menos servidores e tolerar a perda nos negócios que ocorreria quando seus recursos de computação não conseguissem acompanhar a demanda. Nenhuma é uma opção atraente.

As nuvens públicas oferecem uma alternativa de aluguel de curto prazo da capacidade de computação. O varejista, por exemplo, poderia rodar VMs na nuvem durante os períodos de pico, além de seus servidores locais. Isso dá ao varejista acesso aos servidores de que precisa quando precisa deles, sem ter que pagar por eles quando não são necessários.

O custo unitário de rodar servidores na nuvem pode ser maior do que o de rodar o servidor equivalente no data center, mas o custo total de uma mistura de servidores locais e de curto prazo na nuvem ainda pode ser significativamente menor do que o custo de comprar ou alugar para capacidade de pico e deixar recursos ociosos.

Modelo de Pagamento pelo Uso

Relacionado ao modelo de aluguel de curto prazo da computação em nuvem é o modelo de pagamento pelo uso. Quando você executa um servidor virtual na nuvem, normalmente pagará por um período mínimo, como 1 minuto, e pagará por segundo a partir daí. O custo unitário por segundo variará dependendo das características do servidor. Servidores com mais CPUs e memória custarão mais do que servidores com menos CPUs e menos memória.

É importante para os engenheiros de nuvem entenderem o modelo de preços do seu provedor de nuvem. É fácil acumular uma grande conta por servidores e armazenamento se você não estiver monitorando seu uso. De fato, alguns clientes de

nuvem descobrem que executar aplicações na nuvem pode ser mais caro do que executá-las localmente.

Alocação de Recursos Elástica

Outro diferenciador chave entre a computação local e a computação em nuvem pública é a capacidade de adicionar e remover recursos de computação e armazenamento com pouco aviso prévio. Na nuvem, você poderia iniciar 20 servidores em questão de minutos. Em um centro de dados local, poderia levar dias ou semanas para fazer o mesmo se hardware adicional precisasse ser provisionado.

Os provedores de nuvem projetam seus centros de dados com extensos recursos de computação, armazenamento e rede. Eles otimizam seu investimento alugando eficientemente esses recursos para os clientes. Com dados suficientes sobre padrões de uso dos clientes, eles podem prever a capacidade necessária para atender à demanda dos clientes. Como eles têm muitos clientes, a variação na demanda de qualquer um cliente tem pouco efeito no uso geral de seus recursos.

Recursos extensos e a capacidade de mover recursos rapidamente entre clientes permitem que provedores de nuvem pública ofereçam alocação de recursos elástica mais eficientemente do que pode ser feito em centros de dados menores.

Serviços Especializados

Serviços especializados são, por sua natureza, não amplamente compreendidos. Muitos desenvolvedores sabem como desenvolver interfaces de usuário ou consultar um banco de dados, mas menos foram expostos aos detalhes do processamento de linguagem natural ou aprendizado de máquina. Grandes empresas podem ter os recursos financeiros para desenvolver expertise interna em áreas como ciência de dados e visão computacional, mas muitas outras não têm.

Ao oferecer serviços especializados, os provedores de nuvem estão trazendo capacidades avançadas para um público mais amplo de desenvolvedores. Assim como investir em grandes quantidades de hardware, os fornecedores de nuvem pública podem investir em serviços especializados e recuperar seus custos e obter lucro porque os serviços especializados são usados por um grande número de clientes.

Resumo

O Google Cloud oferece uma variedade de serviços para computação, armazenamento, rede e serviços especializados. Serviços de computação incluem máquinas virtuais e clusters Kubernetes, enquanto os serviços de armazenamento suportam armazenamento de objetos e arquivos, junto com caching. Serviços de rede fornecem nuvens privadas virtuais e outros serviços, incluindo VPNs, Interconexões, VPC Compartilhada, emparelhamento de redes VPC e emparelhamento Direto ou por Operadora. Serviços especializados incluem aprendizado de máquina, conversão de fala em texto e serviços de recomendação.

A computação em nuvem tem várias vantagens sobre a computação local, incluindo: alugar em vez de possuir infraestrutura, um modelo de pagamento conforme o uso, alocação elástica de recursos e serviços especializados

Essenciais do Exame

Entenda as diferentes maneiras de fornecer recursos de computação em nuvem. Os recursos computacionais podem ser alocados como VMs individuais ou clusters de VMs que você gerencia. Você também pode usar clusters Kubernetes gerenciados que aliviam parte do overhead operacional de gerenciar um cluster Kubernetes. Opções de computação sem servidor aliviam os usuários de qualquer gerenciamento de servidor. Em vez disso, os desenvolvedores executam seu código em um ambiente containerizado gerenciado pelo provedor de nuvem ou em uma plataforma de computação projetada para código de curta duração. Desenvolvedores e profissionais de DevOps têm o maior controle sobre os recursos quando gerenciam seus próprios servidores e clusters. Serviços gerenciados e opções sem servidor são boas escolhas quando você não precisa de controle sobre o ambiente de computação e obterá mais valor por não ter que gerenciar recursos de computação.

Entenda as diferentes formas de armazenamento em nuvem e quando usá-las. Existem quatro categorias principais de armazenamento: objeto, arquivo, bloco e caches em memória. O armazenamento de objetos é projetado para armazenamento altamente confiável e durável de objetos, como imagens ou conjuntos de dados.

Questões

1. Qual das seguintes opções é uma escolha para selecionar um recurso de computação no Google Cloud?
 - A. Cache
 - B. Máquina virtual (VM)
 - C. Bloco
 - D. Sub-rede
2. Se você usa um cluster gerenciado por um provedor de nuvem, qual destes será gerenciado para você pelo provedor de nuvem?
 - A. Monitoramento
 - B. Rede
 - C. Algumas tarefas de gerenciamento de segurança
 - D. Todas as opções acima
3. Você precisa de computação sem servidor para processamento de arquivos e execução do back end de um site; quais dois produtos você pode escolher do Google Cloud?
 - A. Kubernetes Engine e Compute Engine
 - B. Cloud Run e Cloud Functions
 - C. Cloud Functions e Compute Engine
 - D. Cloud Functions e Kubernetes Engine
4. Foi solicitado a você projetar um sistema de armazenamento para uma aplicação web que permite aos usuários carregar grandes arquivos de dados para serem analisados por um fluxo de trabalho de análise de dados. Os arquivos devem ser armazenados em um sistema de armazenamento de alta disponibilidade. A funcionalidade de sistema de arquivos não é necessária. Qual sistema de armazenamento no Google Cloud deve ser usado?
 - A. Armazenamento de blocos
 - B. Armazenamento de objetos
 - C. Cache
 - D. Sistema de Arquivos de Rede
5. Todos os sistemas de armazenamento de blocos usam qual tamanho de bloco?
 - A. 4 KB.

- B. 8 KB.
 - C. 16 KB.
 - D. O tamanho do bloco pode variar.
6. Você foi solicitado a configurar a segurança de rede em uma nuvem privada virtual. Sua empresa quer ter várias sub-redes e limitar o tráfego entre as sub-redes. Qual controle de segurança de rede você usaria para controlar o fluxo de tráfego entre as sub-redes?
- A. Gerenciamento de acesso de identidade
 - B. Roteador
 - C. Firewall
 - D. Tabela de endereços IP
7. Quando você cria um serviço de aprendizado de máquina para aprender a classificar objetos usando dados tabulares, que tipo de servidores você deve escolher para gerenciar recursos computacionais?
- A. Máquinas virtuais (VMs).
 - B. Clusters de VMs.
 - C. Sem servidores; você deve usar serviços especializados, que são sem servidor.
 - D. VMs executando apenas Linux.
8. Quando é vantajoso investir em servidores por períodos prolongados, como comprometer-se a usar servidores por três a cinco anos?
- A. Quando uma empresa está começando
 - B. Quando uma empresa pode prever com precisão a necessidade de servidores por um período prolongado
 - C. Quando uma empresa tem um orçamento de TI fixo
 - D. Quando uma empresa tem um orçamento de TI variável
9. Sua empresa está baseada na América do Norte e estará executando um servidor virtual para processamento em lote de faturas. Qual fator determina o custo unitário por minuto?
- A. O horário do dia em que a máquina virtual (VM) é executada
 - B. As características do servidor
 - C. O aplicativo que você executa
 - D. Nenhuma das opções acima

10. Você planeja usar o AutoML para analisar dados de vendas e prever a demanda do produto no futuro próximo. Você planeja analisar entre 1.000 e 2.500 produtos por hora. Quantas VMs você deve alocar para atender à demanda de pico?
- A. 1.
 - B. 10.
 - C. 25.
 - D. Nenhuma; AutoML é um serviço sem servidor.
11. Você precisa executar uma série de serviços para suportar uma aplicação. Qual dos seguintes é um bom modelo de implantação?
- A. Executar em uma grande VM única.
 - B. Usar contêineres em um cluster gerenciado.
 - C. Usar duas grandes VMs, tornando uma delas somente leitura.
 - D. Usar uma pequena VM para todos os serviços e aumentar o tamanho da VM quando a utilização da CPU exceder 90%.
12. Você criou uma VM. Quais das seguintes operações de administração de sistema você está autorizado a realizar nela?
- A. Configurar o sistema de arquivos.
 - B. Aplicar patches no software do sistema operacional.
 - C. Alterar permissões de arquivos e diretórios.
 - D. Todas as opções acima.
13. O Cloud Filestore é baseado em qual tecnologia de sistema de arquivos?
- A. Network File System (NFS)
 - B. XFS
 - C. EXT4
 - D. ReiserFS
14. Ao criar recursos no Google Cloud, esses recursos estão sempre parte de quê?
- A. Nuvem privada virtual
 - B. Subdomínio
 - C. Cluster
 - D. Nenhuma das opções acima
15. Você precisa armazenar dados para uma aplicação e está usando um cache. Como o cache afetará a recuperação de dados?
- A. Um cache melhora a execução do JavaScript do lado do cliente.

B. Um cache continuará a armazenar dados mesmo se a energia for perdida, melhorando a disponibilidade.

C. Caches podem ficar desincronizados com o sistema de verdade.

D. Usar um cache reduzirá a latência, já que recuperar de um cache é mais rápido do que recuperar de SSDs ou HDDs.

16. Por que os provedores de nuvem podem oferecer alocação de recursos elástica?

A. Provedores de nuvem podem tirar recursos de clientes de menor prioridade e dá-los a clientes de maior prioridade.

B. Recursos extensivos e a capacidade de mudar rapidamente recursos entre clientes permitem que provedores de nuvem pública ofereçam alocação de recursos elástica de forma mais eficiente do que pode ser feito em data centers menores.

C. Eles cobram mais quanto mais recursos você usa.

D. Eles não oferecem.

17. Qual não é uma característica dos serviços especializados no Google Cloud?

A. São sem servidor; você não precisa configurar servidores ou clusters.

B. Eles fornecem uma função específica, como traduzir texto ou analisar imagens.

C. Eles exigem monitoramento pelo usuário.

D. Eles fornecem uma API para acessar a funcionalidade do serviço.

18. As transações do seu cliente devem acessar um drive anexado a uma VM que permite acesso aleatório a partes dos arquivos. Que tipo de armazenamento o drive anexado fornece?

A. Armazenamento de objetos

B. Armazenamento em blocos

C. Armazenamento NoSQL

D. Armazenamento SQL

19. Você está implantando um novo banco de dados relacional para suportar uma aplicação web. Qual tipo de sistema de armazenamento você usaria para armazenar os arquivos de dados do banco de dados?

A. Armazenamento de objetos

B. Armazenamento de dados

C. Armazenamento em blocos

D. Cache

20. Um usuário prefere serviços que requerem configuração mínima; por que você recomendaria Cloud Storage, Cloud Run e Cloud Functions?

- A. Eles são cobrados apenas pelo tempo.
- B. Eles são sem servidor.
- C. Eles exigem que um usuário configure VMs.
- D. Eles só podem executar aplicações escritas em Go.

Capítulo 2

Serviços de Computação em Nuvem do Google

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO DO ENGENHEIRO ASSOCIADO À NUVEM DO GOOGLE:

✓✓ 2.2 Planejamento e configuração de recursos de computação

✓✓ 3.4 Implantação e implementação de soluções de dados

O Google Cloud é composto por uma vasta gama de serviços que atendem a uma variedade de necessidades de computação, armazenamento e rede. Este capítulo oferece uma visão geral dos serviços de computação em nuvem do Google mais importantes e descreve alguns casos de uso importantes para esses serviços.

Componentes de Computação do Google Cloud

O Google Cloud é um conjunto de serviços de computação em nuvem que inclui serviços de computação, armazenamento e rede projetados para atender às necessidades de uma ampla gama de clientes de computação em nuvem.

Pequenas empresas podem ser atraídas por máquinas virtuais (VMs) e serviços de armazenamento. Grandes empresas e outras organizações de grande porte podem estar mais interessadas em acesso a clusters altamente escaláveis de VMs, uma variedade de bancos de dados relacionais e NoSQL, serviços de rede especializados e capacidades avançadas de inteligência artificial e aprendizado de máquina.

Este capítulo oferece uma visão geral de muitos dos serviços do Google Cloud. A amplitude de serviços disponíveis no Google Cloud continua a crescer. Até o momento em que você ler isso, o Google pode estar oferecendo serviços adicionais. A maioria dos serviços pode ser agrupada em várias categorias principais.

- Recursos de computação
- Recursos de armazenamento
- Bancos de dados
- Serviços de rede
- Gerenciamento de identidade e segurança
- Ferramentas de desenvolvimento
- Ferramentas de gerenciamento
- Serviços especializados

Um Engenheiro Associado à Nuvem certificado pelo Google deve estar familiarizado com os serviços em cada categoria, como eles são usados e as vantagens e desvantagens dos vários serviços em cada categoria.

Recursos de Computação

Os serviços públicos de nuvem oferecem uma gama de opções de serviço de computação. Em um extremo do espectro, os clientes podem criar e gerenciar as VMs por conta própria. Esse modelo dá ao usuário da nuvem o maior controle de todos os serviços de computação. Os usuários podem escolher o sistema operacional a ser executado, quais pacotes instalar e quando fazer backup e realizar outras operações de manutenção. Esse tipo de serviço de computação é tipicamente referido como infraestrutura como serviço (IaaS).

Um modelo alternativo é chamado de plataforma como serviço (PaaS), que fornece um ambiente de execução para executar aplicações sem a necessidade de gerenciar servidores, redes e sistemas de armazenamento subjacentes.

Um dos produtos de computação IaaS é chamado Compute Engine, e as ofertas de PaaS são App Engine e Cloud Functions. Além disso, o Google oferece o Kubernetes

Engine, que é um serviço para gerenciar contêineres em um cluster; esse tipo de serviço é uma alternativa cada vez mais popular para gerenciar conjuntos individuais de VMs.

Compute Engine

O Compute Engine é um serviço que permite aos usuários criar VMs, anexar armazenamento persistente a essas VMs e utilizar outros serviços do Google Cloud, como o Cloud Storage.

VMs são abstrações de servidores físicos. Elas são essencialmente programas que emulam servidores físicos e fornecem CPU, memória, armazenamento e outros serviços que você encontraria se executasse seu sistema operacional favorito em um servidor sob sua mesa ou em um data center. VMs rodam dentro de um serviço de baixo nível chamado hipervisor. O Google Cloud usa uma versão endurecida de segurança do hipervisor KVM. KVM significa Kernel Virtual Machine e fornece virtualização em sistemas Linux rodando em hardware x86.

Hipervisores executam sistemas operacionais como Linux ou Windows Server.

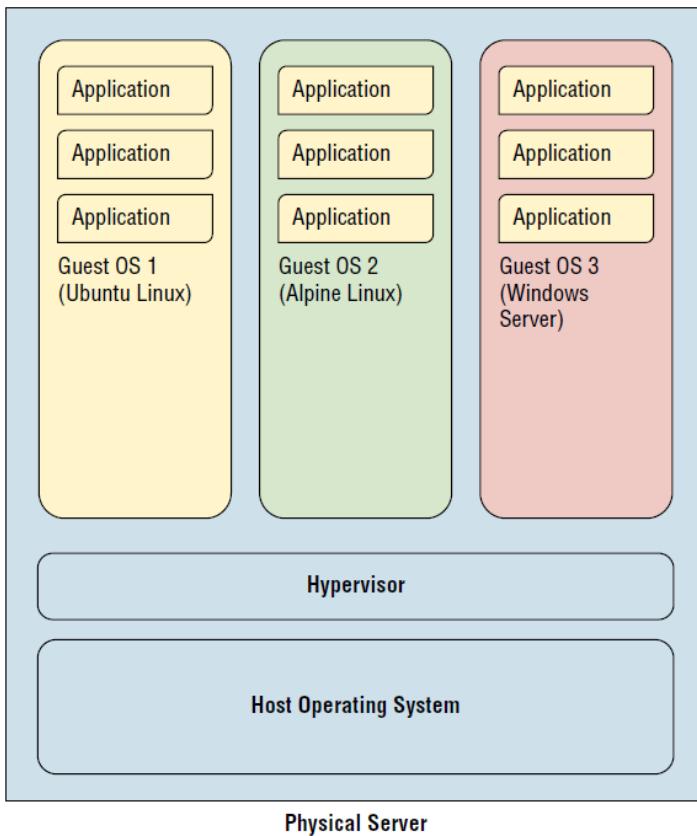
Hipervisores podem rodar múltiplos sistemas operacionais, referidos como sistemas operacionais convidados, enquanto mantém as atividades de cada um isoladas dos outros sistemas operacionais convidados. Cada instância de um sistema operacional convidado executando é uma instância de VM. A Figura 2.1 mostra a organização lógica de instâncias de VM rodando em um servidor físico.

VMs vêm em uma variedade de tamanhos pré-definidos, mas você também pode criar uma configuração personalizada. Ao criar uma instância, você pode especificar vários parâmetros, incluindo os seguintes:

- O sistema operacional
- O tamanho do armazenamento persistente
- Se você adicionará unidades de processamento gráfico (GPUs) para operações intensivas de computação como aprendizado de máquina
- Se você tornará a VM preemptiva

A última opção, tornar uma VM preemptiva, significa que você pode ser cobrado significativamente menos pela VM do que o normal (cerca de 80% menos), mas sua VM pode ser desligada a qualquer momento pelo Google. Ela será desligada após a VM preemptiva ter rodado por pelo menos 24 horas.

FIGURA 2.1 Instâncias de VM rodando dentro de um hipervisor



O Capítulo 4, "Introdução à Computação no Google Cloud", introduzirá os detalhes do gerenciamento de VMs do Compute Engine. Para explorar o Compute Engine, faça login no Console do Google Cloud, navegue até o menu principal à esquerda e selecione Compute Engine.

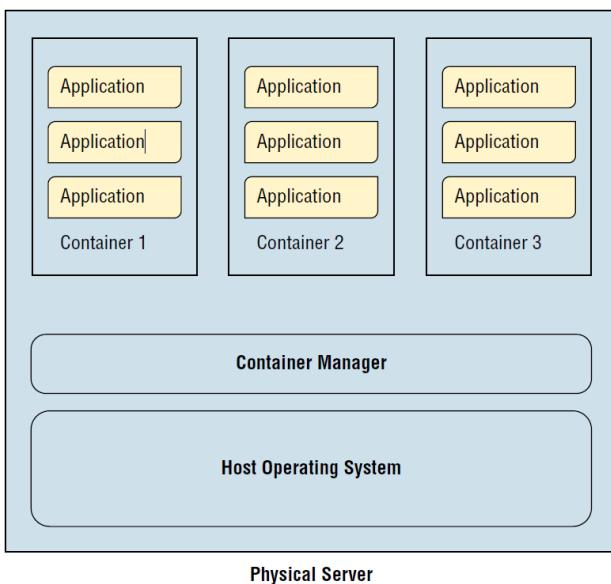
Kubernetes Engine

O Kubernetes Engine é projetado para permitir que os usuários executem facilmente aplicações contêinerizadas em um cluster de servidores. Os contêineres são frequentemente comparados às VMs porque ambos são usados para isolar processamento de computação e recursos. Contêineres adotam uma abordagem diferente das VMs para isolar processos de computação.

Como mencionado, uma VM executa um sistema operacional convidado em um servidor físico. O servidor físico também executa um sistema operacional, junto com um hipervisor. Outra abordagem para isolar recursos de computação é usar recursos do sistema operacional hospedeiro para isolar processos e recursos. Com essa abordagem, não há necessidade de um hipervisor; o sistema operacional hospedeiro mantém o isolamento. Em vez disso, um gerenciador de contêineres é usado. Ou seja, um único gerenciador de contêiner coordena os contêineres que estão sendo executados no servidor.

Não há sistemas operacionais adicionais, ou convidados, executando em cima do gerenciador de contêineres. Em vez disso, os contêineres utilizam a funcionalidade do sistema operacional hospedeiro, enquanto o sistema operacional e o gerenciador de contêineres garantem o isolamento entre os contêineres em execução. A Figura 2.2 mostra a estrutura lógica dos contêineres.

FIGURA 2.2 Contêineres rodando em um servidor físico



O Kubernetes Engine é um produto do Google Cloud que permite aos usuários descrever os recursos de computação, armazenamento e memória que gostariam de utilizar para executar seus serviços. O Kubernetes Engine, então, provisão os recursos subjacentes. É fácil adicionar e remover recursos de um cluster Kubernetes usando uma interface de linha de comando ou uma interface gráfica do usuário.

Além disso, o Kubernetes monitora a saúde dos servidores no cluster e repara automaticamente problemas, como servidores que falharam. O Kubernetes Engine também suporta o escalonamento automático, então, se a carga em suas aplicações aumentar, o Kubernetes Engine alocará recursos adicionais.

Os clusters Anthos estendem o GKE para ambientes híbridos e multicloud, fornecendo serviços para criar, dimensionar e atualizar clusters Kubernetes conformes, juntamente com uma camada de orquestração comum. Vários clusters podem ser gerenciados como um grupo conhecido como frota. Os clusters Anthos podem ser conectados usando opções de rede padrão, incluindo VPNs, Dedicated Interconnect e Partner Interconnects.

Existem vários benefícios chave em usar o Anthos para gerenciar múltiplos clusters Kubernetes. Estes incluem:

- Gerenciamento centralizado de configuração como código
- Capacidade de reverter implantações com o Git
- Uma visão única da infraestrutura de cluster e aplicações
- Fluxos de trabalho centralizados e auditáveis
- Instrumentação de código usando o Anthos Service Mesh
- Autorização e roteamento do Anthos Service Mesh

Além disso, o Anthos inclui o Migrate for Anthos para GKE, que é um serviço que permite orquestrar migrações usando Kubernetes e Anthos.

O termo “clusters Anthos” refere-se aos clusters do Google Kubernetes Engine que foram estendidos para funcionar em ambientes locais ou em multicloud.

O Capítulo 7, “Computação com Kubernetes,” descreverá os detalhes do planejamento e gerenciamento do Kubernetes Engine. Para explorar o Kubernetes Engine, faça login no Console do Google Cloud, navegue até o menu principal à esquerda e selecione Kubernetes Engine.

App Engine

O App Engine é uma oferta PaaS de computação do Google Cloud. Com o App Engine, desenvolvedores e administradores de aplicações não precisam se preocupar com a configuração de VMs ou especificação de clusters Kubernetes. Em vez disso, os desenvolvedores criam aplicações em uma linguagem de programação popular, como Java, Go, Python ou Node.js, e implantam esse código em um ambiente de aplicação sem servidor.

O App Engine gerencia a infraestrutura de computação e rede subjacente. Não há necessidade de configurar VMs ou proteger redes para proteger sua aplicação. O App Engine é bem adequado para aplicações back-end de web e mobile.

O App Engine está disponível em dois tipos:

■■ No ambiente padrão, você executa aplicações em uma sandbox específica da linguagem, então sua aplicação é isolada do sistema operacional do servidor subjacente, bem como de outras aplicações que estão sendo executadas nesse servidor. O ambiente padrão é bem adequado para aplicações escritas em uma das linguagens suportadas e que não precisam de pacotes do sistema operacional ou outro software compilado que teria que ser instalado junto com o código da aplicação.

■■ No ambiente flexível, você executa aplicações contêinerizadas no ambiente do App Engine. O ambiente flexível funciona bem nos casos em que você tem código de aplicação, mas também precisa de bibliotecas ou outro software de terceiros instalado. Como o nome sugere, o ambiente flexível oferece mais opções, incluindo a capacidade de trabalhar com processos em segundo plano e escrever em disco local.

O Capítulo 9, “Computando com Cloud Run e App Engine,” introduzirá detalhes para usar e gerenciar o App Engine. Para explorar o App Engine, faça login no Console do Google Cloud, navegue até o menu principal à esquerda e selecione App Engine.

Cloud Run

O Cloud Run é um serviço do Google Cloud para executar contêineres sem estado. Ao usar o serviço gerenciado, você paga por uso e pode ter até 1.000 instâncias de contêiner por padrão. Ao contrário do App Engine Standard, o Cloud Run não o restringe a usar um conjunto fixo de linguagens de programação. Os serviços Cloud Run têm disponibilidade regional.

Um serviço é a principal abstração de computação no Cloud Run. Um serviço está em uma região e replicado em várias zonas. Um serviço pode ter várias revisões. O Cloud Run irá escalar automaticamente o número de instâncias com base na carga.

Cloud Functions

O Google Cloud Functions é uma opção de computação leve que é bem adequada para processamento baseado em eventos. Cloud Functions executa código em resposta a um evento, como um arquivo sendo carregado para o Cloud Storage ou uma mensagem sendo escrita em uma fila de mensagens. O código que executa no ambiente do Cloud Functions deve ser de curta duração — este serviço de computação não é projetado para executar código de longa duração. Se você precisa suportar aplicações ou trabalhos de longa duração, considere o Compute Engine, Kubernetes Engine ou App Engine.

Cloud Functions é frequentemente usado para chamar outros serviços, como interfaces de programação de aplicações (APIs) de terceiros ou outros serviços do Google Cloud, como um serviço de tradução de linguagem natural.

Como o App Engine e o Cloud Run, Cloud Functions é um produto sem servidor (serverless). Os usuários só precisam fornecer o código; eles não precisam configurar VMs ou criar contêineres. O Cloud Functions escalará automaticamente à medida que a carga aumentar.

Além desses produtos de computação, o Google Cloud oferece vários recursos de armazenamento.

O Capítulo 10, "Computando com Cloud Functions", descreverá os detalhes de uso e gerenciamento do Cloud Functions. Para explorar o Cloud Functions, faça login no Console do Google Cloud, navegue até o menu principal à esquerda e selecione Cloud Functions.

Componentes de Armazenamento do Google Cloud

Aplicações e serviços que rodam na nuvem devem atender a uma ampla gama de requisitos quando se trata de armazenamento.

Recursos de Armazenamento

Às vezes, uma aplicação precisa de tempos rápidos de leitura e escrita para quantidades moderadas de dados. Outras vezes, uma aplicação empresarial pode precisar de acesso a petabytes de armazenamento arquivístico, mas pode tolerar minutos e até horas para recuperar um documento. O Google Cloud possui vários recursos de armazenamento para armazenar objetos e arquivos.

Cloud Storage

O Cloud Storage é o sistema de armazenamento de objetos do Google Cloud. Os objetos podem ser qualquer tipo de arquivo ou objetos grandes binários, conhecidos como blob. Os objetos são organizados em buckets, que são análogos a diretórios em um sistema de arquivos. É importante lembrar que o Cloud Storage não é um sistema de arquivos; é um serviço que recebe, armazena e recupera arquivos ou objetos de um sistema de

armazenamento distribuído. O Cloud Storage não faz parte de uma VM da mesma forma que um disco persistente anexado. O Cloud Storage é acessível a partir de VMs, contêineres ou qualquer outro dispositivo de rede com privilégios apropriados e, portanto, complementa os sistemas de arquivos em discos persistentes.

Cada objeto armazenado é endereçável de forma única por uma URL. Por exemplo, uma versão em PDF deste capítulo, chamada chapter1.pdf, que se armazenada em um bucket chamado ace-certification-exam-prep, seria endereçável da seguinte forma:

<https://storage.cloud.google.com/ace-certification-exam-prep/chapter1.pdf>

Usuários do Google Cloud e outros podem ser autorizados a ler e escrever objetos em um bucket. Muitas vezes, uma aplicação será concedida privilégios por meio de uma conta de serviço com papéis de Gerenciamento de Identidade e Acesso (IAM) para habilitar a aplicação a ler e escrever em buckets.

O Cloud Storage é útil para armazenar objetos que são tratados como unidades únicas de dados. Por exemplo, um arquivo de imagem é um bom candidato para armazenamento de objetos. Imagens geralmente são lidas e escritas de uma só vez. Raramente é necessário recuperar apenas uma parte da imagem. Em geral, se você escrever ou recuperar um objeto de uma vez e precisar armazená-lo independentemente de servidores que podem ou não estar em execução a qualquer momento, então o Cloud Storage é uma boa opção.

Existem diferentes tipos de localização para armazenamento em nuvem. O armazenamento regional mantém cópias de objetos em uma única região do Google Cloud. Regiões são áreas geográficas distintas que podem ter várias zonas, ou áreas de implantação. Uma zona é considerada um único domínio de falha, o que significa que se todas as instâncias de sua aplicação estiverem em execução em uma zona e houver uma falha, todas as instâncias de sua aplicação ficarão inacessíveis. O armazenamento regional é bem adequado para aplicações que são executadas na mesma região e precisam de acesso de baixa latência aos objetos no Cloud Storage.

O Cloud Storage tem alguns recursos avançados úteis, como suporte para várias regiões. Esse recurso permite armazenar réplicas de objetos em várias regiões do Google Cloud, o que é importante para alta disponibilidade, durabilidade e baixa latência.

Cenário Real

Armazenamento Multi-Região

Se houvesse uma interrupção na região us-east1 e seus objetos estivessem armazenados apenas nessa região, você não poderia acessar esses objetos durante a interrupção. No entanto, se você habilitasse o armazenamento multi-região, então seus objetos armazenados em us-east1 também estariam armazenados em outra região, como us-west1.

Além de alta disponibilidade e durabilidade, o armazenamento multi-região permite um acesso mais rápido aos dados quando usuários ou aplicações estão distribuídos pelas regiões.

Às vezes, os dados precisam ser mantidos por períodos prolongados, mas raramente são acessados. Nesses casos, as classes de armazenamento nearline e coldline são boas opções. Use nearline quando você acessará objetos menos de uma vez por mês, e use coldline quando você acessará objetos menos de uma vez a cada 90 dias.

A classe de armazenamento arquivo é um armazenamento arquivístico de baixo custo projetado para alta durabilidade e acesso infrequente. Esta classe de armazenamento é adequada para dados que são acessados menos de uma vez por ano.

Um recurso útil do Cloud Storage é o conjunto de políticas de gerenciamento de ciclo de vida que podem gerenciar automaticamente objetos com base em políticas que você define. Por exemplo, você poderia definir uma política que move todos os objetos com mais de 60 dias em um bucket da classe de armazenamento padrão para a classe de armazenamento nearline, ou exclui qualquer objeto em um bucket de armazenamento de arquivo que seja mais antigo do que cinco anos.

Disco Persistente

Discos persistentes são serviços de armazenamento que são anexados a VMs no Compute Engine ou Kubernetes Engine. Discos persistentes fornecem armazenamento de blocos em unidades de estado sólido (SSDs) e discos rígidos (HDDs). SSDs são frequentemente usados para aplicações de baixa latência onde o desempenho do disco persistente é importante. SSDs custam mais do que HDDs, então aplicações que requerem grandes quantidades de armazenamento de disco persistente, mas podem tolerar tempos de leitura e escrita mais longos, podem usar HDDs para atender às suas necessidades de armazenamento.

Uma vantagem dos discos persistentes no Google Cloud é que esses discos suportam múltiplos leitores sem degradação no desempenho. Isso permite que múltiplas instâncias leiam uma única cópia dos dados. Os discos também podem ser redimensionados conforme necessário enquanto estão em uso, sem a necessidade de reiniciar suas VMs.

Discos persistentes podem ter até 64 TB de tamanho, usando tanto SSDs quanto HDDs. Vários discos persistentes podem ser anexados a uma única VM.

Armazenamento em Nuvem para Firebase

Desenvolvedores de aplicativos móveis podem achar o Armazenamento em Nuvem para Firebase a melhor combinação de armazenamento de objetos na nuvem e a capacidade de suportar uploads e downloads de dispositivos móveis com conexões de rede às vezes instáveis.

A API do Armazenamento em Nuvem para Firebase é projetada para fornecer transmissão segura, bem como mecanismos de recuperação robustos para lidar com a qualidade de rede potencialmente problemática. Uma vez que arquivos, como fotos ou gravações de música, são carregados no Armazenamento em Nuvem, você pode acessar esses arquivos através da interface de linha de comando do Armazenamento em Nuvem e kits de desenvolvimento de software (SDKs).

Cloud Filestore

Às vezes, desenvolvedores precisam ter acesso a um sistema de arquivos alojado em armazenamento conectado à rede. Para esses casos de uso, o serviço Cloud Filestore oferece um sistema de arquivos compartilhado para uso com o Compute Engine e o Kubernetes Engine.

O Filestore pode fornecer um alto número de operações de entrada e saída por segundo (IOPS) bem como capacidade de armazenamento variável. Administradores de sistema de arquivos podem configurar o Cloud Filestore para atender às suas necessidades específicas de IOPS e capacidade.

O Filestore implementa o protocolo Network File System (NFS) para que administradores de sistema possam montar facilmente sistemas de arquivos compartilhados em servidores virtuais.

Sistemas de armazenamento como os descritos são usados para armazenar objetos de grãos grossos, como arquivos. Quando os dados são mais finamente estruturados e precisam ser recuperados usando linguagens de consulta que descrevem o subconjunto de dados a ser retornado, então é melhor usar um sistema de gerenciamento de banco de dados.

O Capítulo 11, "Planejando Armazenamento na Nuvem", descreve detalhes e orientações para o planejamento de serviços de armazenamento. Para explorar opções de armazenamento, faça login no Console do Google Cloud, navegue até o menu principal à esquerda e selecione Armazenamento ou Filestore.

Bancos de Dados

O Google Cloud oferece várias opções de banco de dados. Alguns são bancos de dados relacionais e outros são bancos de dados NoSQL. Alguns são sem servidor (serverless) e outros requerem que os usuários gerenciem clusters de servidores. Alguns oferecem suporte para transações atômicas, e outros são mais adequados para aplicações com requisitos de consistência e transação menos rigorosos. Os usuários do Google Cloud devem entender os requisitos de suas aplicações antes de escolher um serviço, e isso é especialmente importante ao escolher um banco de dados, que frequentemente fornece serviços de armazenamento central na pilha da aplicação.

O Cloud SQL é um serviço gerenciado de banco de dados relacional do Google Cloud que permite aos usuários configurar bancos de dados MySQL, PostgreSQL e SQL Server em VMs sem ter que se preocupar com tarefas de administração de banco de dados, como fazer backup de bancos de dados ou aplicar patches no software de banco de dados. Este serviço de banco de dados inclui gerenciamento de replicação e permite failover automático, proporcionando bancos de dados altamente disponíveis.

Bancos de dados relacionais são bem adequados para aplicações com requisitos de estrutura de dados relativamente consistentes. Por exemplo, um banco de dados bancário pode rastrear números de contas, nomes de clientes, endereços, etc. Já que praticamente todos os registros no banco de dados precisarão das mesmas informações, essa aplicação é uma boa escolha para um banco de dados relacional.

Cloud Bigtable

O Cloud Bigtable é projetado para aplicações de escala de petabytes que podem gerenciar até bilhões de linhas e milhares de colunas. Baseia-se em um modelo NoSQL conhecido como modelo de dados de coluna larga, que é diferente dos bancos de dados relacionais como o Cloud SQL. O Bigtable é adequado para aplicações que requerem operações de escrita e leitura de baixa latência. É projetado para suportar milhões de operações por segundo. O Bigtable integra-se com outros serviços do Google Cloud, como Cloud Storage, Cloud Pub/Sub, Cloud Dataflow e Cloud Dataproc. Também suporta a API HBase, que é uma API para acesso a dados no ecossistema de big data do Hadoop. O Bigtable também se integra com ferramentas de código aberto para processamento de dados, análise de grafos e análise de séries temporais.

Cloud Spanner

O Cloud Spanner é o banco de dados relacional globalmente distribuído do Google que combina os principais benefícios dos bancos de dados relacionais, como consistência forte e transações, com a capacidade de escalar horizontalmente como um banco de dados NoSQL. O Spanner é um banco de dados de alta disponibilidade com um acordo de nível de serviço (SLA) de 99,999% de disponibilidade, tornando-o uma boa opção para aplicações empresariais que exigem serviços de banco de dados relacionais escaláveis e altamente disponíveis. O Cloud Spanner também possui segurança de nível empresarial com criptografia em repouso e em trânsito, juntamente com controles de acesso baseados em identidade. O Cloud Spanner suporta SQL padrão ANSI 2011.

Cloud Firestore

O Cloud Firestore, anteriormente conhecido como Cloud Datastore, é um banco de dados de documentos NoSQL. Esse tipo de banco de dados usa o conceito de um documento, ou coleção de pares chave-valor, como o bloco de construção básico. Os documentos permitem esquemas flexíveis. Por exemplo, um documento sobre um livro pode ter pares chave-valor listando autor, título e data de publicação. Alguns livros também podem ter informações sobre sites companheiros e traduções para outros idiomas. O conjunto de chaves que podem ser incluídas não precisa ser definido antes do uso em bancos de dados de documentos. Isso é especialmente útil quando aplicações devem acomodar uma gama de atributos, alguns dos quais podem não ser conhecidos no momento do design. O Cloud Firestore é acessado via uma API REST que pode ser usada a partir de aplicações rodando no Compute Engine, Kubernetes Engine ou App Engine. Este banco de dados escalará automaticamente baseado na carga. Ele também dividirá, ou particionará, dados conforme necessário para manter o desempenho. Como o Cloud Firestore é um serviço gerenciado, ele cuida de replicação, backups e outras tarefas de administração de banco de dados.

Embora seja um banco de dados NoSQL, o Cloud Firestore suporta transações, índices e consultas semelhantes ao SQL. O Cloud Firestore é bem adequado para aplicações que exigem alta escalabilidade e dados estruturados e que nem sempre precisam de consistência forte ao ler dados. Catálogos de produtos, perfis de usuários e

histórico de navegação do usuário são exemplos de tipos de aplicações que usam o Cloud Datastore.

Cloud Memorystore

O Cloud Memorystore é um serviço de cache em memória. Outros bancos de dados oferecidos no Google Cloud são projetados para armazenar grandes volumes de dados e suportar consultas complexas, mas o Cloud Memorystore é um serviço gerenciado para fazer cache de dados frequentemente usados em memória. Caches como este são usados para reduzir o tempo necessário para ler dados em uma aplicação. O Cloud Memorystore é projetado para fornecer acesso aos dados em submilissegundos. O Cloud Memorystore suporta tanto Redis quanto memcached, dois populares sistemas de cache de código aberto.

Como um serviço gerenciado, o Cloud Memorystore permite aos usuários especificar o tamanho de um cache enquanto deixa tarefas de administração para o Google. O Google Cloud garante alta disponibilidade, aplicação de patches e failover automático para que os usuários não tenham que se preocupar.

O Capítulo 12, “Implantando Armazenamento no Google Cloud”, aprofunda-se nos detalhes de como criar vários tipos de bancos de dados, bem como como carregar, excluir e consultar dados. Cada um dos bancos de dados pode ser acessado a partir do menu principal do Console do Google Cloud. A partir daí, você pode começar a explorar como cada um funciona e começar a ver as diferenças.

Componentes de Rede do Google Cloud

Nesta seção, revisaremos os principais componentes de rede. Detalhes sobre a configuração de redes e sua gestão são descritos no Capítulo 14, “Redes na Nuvem: Nuvens Privadas Virtuais e Redes Privadas Virtuais”, e no Capítulo 15, “Redes na Nuvem: DNS, Balanceamento de Carga, Acesso Privado do Google e Endereçamento IP”.

Serviços de Rede

O Google Cloud oferece vários serviços de rede projetados para permitir que os usuários configurem redes virtuais dentro da infraestrutura de rede global do Google, liguem data centers locais à rede do Google, otimizem a entrega de conteúdo e protejam seus recursos na nuvem usando serviços de segurança de rede.

Nuvem Privada Virtual

Quando uma empresa opera seu próprio data center, ela controla o que está fisicamente localizado nesse data center e conectado à sua rede. Sua infraestrutura é fisicamente isolada daquelas de outras organizações que operam em outros data centers. Quando uma organização se muda para uma nuvem pública, ela compartilha infraestrutura com outros clientes dessa nuvem pública. Embora várias empresas usem a mesma infraestrutura de nuvem, cada empresa pode isolar logicamente seus recursos na nuvem criando uma nuvem privada virtual (VPC).

Uma característica distintiva do Google Cloud é que uma VPC pode abranger o globo sem depender da Internet pública. O tráfego de qualquer servidor em uma VPC pode ser roteado com segurança pela rede global do Google para qualquer outro ponto nessa rede. Outra vantagem da estrutura de rede do Google é que seus servidores de back-end podem acessar serviços do Google, como aprendizado de máquina ou serviços da Internet das Coisas (IoT), sem criar um endereço IP público para servidores de back-end.

VPCs no Google Cloud podem ser ligadas a redes privadas virtuais locais usando Segurança de Protocolo de Internet (IPSec).

Embora uma VPC seja global, as empresas podem usar projetos e contas de cobrança separadas para gerenciar diferentes departamentos ou grupos dentro da organização. Firewalls também podem ser usados para restringir o acesso a recursos em uma VPC.

Balanceamento de Carga na Nuvem

O Google oferece balanceamento de carga global para distribuir cargas de trabalho em sua infraestrutura na nuvem. Usando um único endereço IP anycast, o Cloud Load Balancing pode distribuir a carga de trabalho dentro e entre regiões, adaptar-se a servidores falhos ou degradados e escalar automaticamente seus recursos de computação para acomodar mudanças na carga de trabalho. O Cloud Load Balancing também suporta balanceamento de carga interno, portanto, não é necessário expor endereços IP à Internet para obter as vantagens do balanceamento de carga.

O Cloud Load Balancing é um serviço de software que pode平衡ear cargas de tráfego HTTP, HTTPS, TCP/SSL e UDP.

Cloud Armor

Serviços expostos à Internet podem se tornar alvos de ataques distribuídos de negação de serviço (DDoS). O Cloud Armor é um serviço de segurança de rede do Google que se baseia no serviço Global HTTP(s) Load Balancing. Os recursos do Cloud Armor incluem o seguinte:

- Capacidade de permitir ou restringir o acesso com base no endereço IP
- Regras predefinidas para combater ataques de cross-site scripting
- Capacidade de combater ataques de injeção SQL
- Capacidade de definir regras do nível 3 (rede) ao nível 7 (aplicação)
- Permite e restringe o acesso com base na geolocalização do tráfego de entrada

Cloud CDN

Com redes de entrega de conteúdo (CDNs), usuários em qualquer lugar podem solicitar conteúdo de sistemas distribuídos em várias regiões. As CDNs permitem resposta de baixa latência a esses pedidos, armazenando o conteúdo em um conjunto de pontos de presença em todo o mundo. Atualmente, o Google tem mais de 100 pontos de presença

de CDN que são gerenciados como um recurso global, portanto, não há necessidade de manter configurações específicas para cada região.

As CDNs são especialmente importantes para sites com grandes quantidades de conteúdo estático e uma audiência global. Sites de notícias, por exemplo, poderiam usar o serviço Cloud CDN para garantir uma resposta rápida a solicitações de qualquer ponto do mundo.

Cloud Interconnect

O Cloud Interconnect é um conjunto de serviços do Google Cloud para conectar suas redes existentes à rede do Google. O Cloud Interconnect oferece dois tipos de conexões: interconexões e peering.

Interconexão com acesso direto a redes usa o padrão Alocação de Endereços para Internets Privadas (RFC 1918) para se conectar a dispositivos na sua VPC. Uma conexão de rede direta é mantida entre um data center local ou hospedado e uma das instalações de colocation do Google, que estão na América do Norte, América do Sul, Europa, Ásia e Austrália. Alternativamente, se uma organização não puder alcançar uma interconexão direta com uma instalação do Google, ela poderia usar o Partner Interconnect. Este serviço depende de um provedor de rede de terceiros para fornecer conectividade entre o data center da empresa e uma instalação do Google.

Parceiro Interconectado é a maneira recomendada de se conectar ao Google Cloud através de provedores, mas se você precisar acessar aplicações do Google Workspace, então você pode usar o emparelhamento de operadoras. O emparelhamento não utiliza recursos do Google Cloud, como conexões de interconect ou Cloud Routers.

Para organizações que não requerem a largura de banda de um interconect direto ou emparelhado, o Google oferece serviços de VPN que permitem o tráfego transmitir entre data centers, outras nuvens de fornecedores e o Google Cloud usando a Internet pública.

Cloud DNS

Cloud DNS é um serviço de nome de domínio fornecido no Google Cloud. Cloud DNS é um serviço de alta disponibilidade e baixa latência para mapear nomes de domínio, como example.com, para endereços IP, como 74.120.28.18.

Cloud DNS é projetado para escalar automaticamente para que os clientes possam ter milhares ou até milhões de endereços sem preocupação com a escala da infraestrutura subjacente. Cloud DNS também oferece zonas privadas que permitem criar nomes personalizados para suas VMs, se necessário.

Gerenciamento de Identidade e Segurança

O serviço de Gerenciamento de Identidade e Acesso (IAM) do Google Cloud permite aos clientes definir controles de acesso refinados em recursos na nuvem. IAM usa os conceitos de usuários, papéis e permissões.

Identidades são abstrações sobre usuários de serviços, como um usuário humano. Depois que uma identidade é autenticada ao fazer login ou por algum outro mecanismo, o usuário autenticado pode acessar recursos e realizar operações baseadas nas permissões concedidas a essa identidade. Por exemplo, um usuário pode ter permissões para criar um bucket no Cloud Storage ou deletar uma VM rodando no Compute Engine.

Usuários frequentemente precisam de conjuntos similares de permissões. Alguém que pode criar uma VM provavelmente também quererá poder modificar ou deletar essas VMs. Grupos de permissões relacionadas podem ser agrupados em papéis. Papéis são conjuntos de permissões que podem ser atribuídos a uma identidade.

Como um Engenheiro Associado Certificado em Nuvem do Google, você se familiarizará com identidades, papéis e permissões e como administrá-los em organizações e projetos.

Você pode encontrar ferramentas de gerenciamento de identidade sob o menu IAM e Admin no Console do Google Cloud. O Capítulo 17, "Configurando Acesso e Segurança", fornece detalhes sobre identidade, papéis e melhores práticas para sua gestão.

Ferramentas de Desenvolvimento

O Google Cloud é uma excelente escolha para desenvolvedores e engenheiros de software devido ao fácil acesso a serviços de infraestrutura e gerenciamento de dados, mas também pelas ferramentas que suporta.

Cloud SDK é uma interface de linha de comando para gerenciar recursos do Google Cloud, incluindo VMs, armazenamento de disco, firewalls de rede e virtualmente qualquer outro recurso que você possa implantar no Google Cloud, além de uma interface de linha de comando, as bibliotecas de clientes do Cloud SDK incluem bibliotecas para Java, Python, Node.js, Ruby, Go, .NET e PHP. O Google Cloud também suporta o deploy de aplicações em contêineres com Container Registry, Cloud Build e Cloud Source Repositories.

O Google também desenvolveu plugins para facilitar o trabalho com ferramentas de desenvolvimento populares. Estes incluem o seguinte:

- Cloud Tools for IntelliJ
- Cloud Tools for PowerShell
- Cloud Tools for Visual Studio
- Cloud Tools for Eclipse
- App Engine Gradle Plugin
- App Engine Maven Plugin

Claro, as aplicações passam do desenvolvimento ao deploy de produção, e o Google Cloud acompanha esse fluxo com ferramentas de gerenciamento adicionais para ajudar a monitorar e manter as aplicações após serem implantadas.

Componentes Adicionais do Google Cloud

As ferramentas de gerenciamento são projetadas para aqueles responsáveis por garantir a confiabilidade, disponibilidade e escalabilidade das aplicações.

Ferramentas de Gerenciamento e Observabilidade

A seguir, estão algumas das ferramentas mais importantes na categoria de ferramentas de gerenciamento e observabilidade:

Cloud Monitoring: Este serviço coleta dados de desempenho do Google Cloud, recursos da AWS e instrumentação de aplicações, incluindo sistemas de código aberto populares como NGINX, Cassandra e Elasticsearch.

Cloud Logging: Este serviço permite aos usuários armazenar, analisar e criar alertas sobre dados de log tanto do Google Cloud quanto de logs da Amazon Web Services (AWS).

Error Reporting: Agrega informações de falhas de aplicações para exibição em uma interface centralizada.

Cloud Trace: É um serviço de rastreamento distribuído que captura dados de latência sobre uma aplicação para ajudar a identificar áreas problemáticas de desempenho.

Depurador na Nuvem (**Cloud Debugger**) Permite que os desenvolvedores inspecionem o estado do código em execução, injetem comandos e visualizem variáveis da pilha de chamadas.

Perfilador na Nuvem (**Cloud Profiler**) É usado para coletar informações sobre a utilização de CPU e memória através da hierarquia de chamadas de uma aplicação. O Profiler usa amostragem estatística para minimizar o impacto da criação de perfis no desempenho da aplicação.

A combinação de ferramentas de gerenciamento e observabilidade fornece insights sobre aplicações enquanto elas rodam em produção, permitindo um monitoramento e análise mais eficazes dos sistemas operacionais.

Serviços Especializados

Além das ofertas de IaaS e PaaS, o Google Cloud possui serviços especializados para APIs, análise de dados e aprendizado de máquina.

Plataforma API Apigee

A plataforma API Apigee é um serviço de gerenciamento para clientes do Google Cloud que fornecem acesso à API para suas aplicações. A plataforma Apigee permite que desenvolvedores implantem, monitorem e protejam suas APIs. Ela também gera proxies de API com base na Especificação de API Aberta.

É difícil prever a carga em uma API, e às vezes podem ocorrer picos de uso. Para esses momentos, a plataforma API Apigee fornece roteamento e limitação de taxa baseada em políticas que os clientes podem definir.

As APIs podem ser autenticadas usando OAuth 2.0 ou SAML. Os dados são criptografados tanto em trânsito quanto em repouso na plataforma API Apigee.

Análise de Dados e Pipelines de Dados

O Google Cloud possui vários serviços projetados para analisar grandes dados em modos batch e streaming. Algumas das ferramentas mais importantes neste conjunto de serviços são:

- BigQuery, um serviço de banco de dados para análise em escala de petabytes para armazenamento de dados
- Cloud Dataflow, uma estrutura para definir pipelines de processamento batch e stream
- Cloud Dataproc, um serviço gerenciado de Hadoop e Spark
- Cloud Dataprep, um serviço que permite aos analistas explorar e preparar dados para análise

Frequentemente, projetos de análise de dados e armazenamento de dados usam vários desses serviços juntos.

IA e Aprendizado de Máquina

O Google é líder em IA e aprendizado de máquina, por isso não é surpresa que o Google Cloud inclua vários serviços de IA. Vertex AI é uma plataforma unificada de IA para construir modelos de aprendizado de máquina. Os serviços especializados nesta área incluem o seguinte:

AutoML É uma ferramenta que permite a desenvolvedores sem experiência em aprendizado de máquina desenvolver modelos de aprendizado de máquina.

IA de Tradução: Essa ferramenta é para traduzir linguagens humanas e inclui AutoML Translation e Translation API para traduções de texto, e Media Translation API para traduções de áudio.

Linguagem Natural: Analisar e extraír características e conceitos do texto usando métodos de aprendizado de máquina.

Visão AI: Esta é uma plataforma de análise de imagem para anotar imagens com metadados, extraír texto ou filtrar conteúdo.

Recomendações AI: Este é um serviço para fornecer recomendações personalizadas aos clientes em grande escala.

Resumo

O Google Cloud oferece uma gama completa de serviços para apoiar o processamento de informações, incluindo recursos de computação, recursos de armazenamento, bancos de dados, serviços de rede, serviços de gerenciamento de identidade e segurança, ferramentas de desenvolvimento, serviços de gerenciamento e operações, bem como serviços especializados para apoiar a IA.

Essenciais do Exame

Entenda as diferenças entre Compute Engine, Kubernetes Engine, App Engine, Cloud Run e Cloud Functions. Compute Engine é o serviço de VM do Google. Os usuários podem escolher CPUs, memória, discos persistentes e sistemas operacionais. Eles podem personalizar ainda mais uma VM adicionando unidades de processamento gráfico para operações intensivas de computação. As VMs são gerenciadas individualmente ou em grupos de servidores semelhantes.

O Kubernetes Engine gerencia grupos de servidores virtuais e aplicações que rodam em contêineres. Contêineres são mais leves do que VMs. Kubernetes é chamado de serviço de orquestração porque distribui contêineres em clusters, monitora a saúde do cluster e escala conforme prescrito pelas configurações.

O App Engine é o PaaS do Google. Desenvolvedores podem rodar seu código em uma sandbox específica da linguagem ao usar o ambiente padrão ou em um contêiner ao usar o ambiente flexível. App Engine é um serviço sem servidor, então os clientes não precisam especificar configurações de VM ou gerenciar servidores.

O Cloud Run é um serviço para rodar contêineres sem estado. Esta é uma opção sem servidor que oferece algumas das vantagens do Kubernetes sem exigir que você implante seus próprios clusters. Observe que o Cloud Run atualmente não suporta aplicações que mantêm estado no contêiner.

O Cloud Functions é um serviço sem servidor projetado para executar códigos de curta duração que respondem a eventos, como uploads de arquivos ou mensagens publicadas em uma fila de mensagens. As funções podem ser escritas em Node.js ou Python.

Entenda o que significa sem servidor. Sem servidor significa que os clientes que usam um serviço não precisam configurar, monitorar ou manter os recursos de computação subjacentes ao serviço. Isso não significa que não haja servidores envolvidos—sempre há servidores físicos que executam aplicações, funções e outros softwares. Sem servidor refere-se apenas à não necessidade de gerenciar esses recursos subjacentes.

Entenda a diferença entre armazenamento de objeto e armazenamento de arquivo. As lojas de objetos são usadas para armazenar e acessar recursos baseados em arquivos. Esses objetos são referenciados por um identificador único, como uma URL. Lojas de objetos não fornecem serviços de bloco ou sistema de arquivos, portanto, não são adequadas para armazenamento de banco de dados. O Cloud Storage é o serviço de armazenamento de objetos do Google Cloud.

O armazenamento de arquivos suporta acesso baseado em bloco a arquivos. Os arquivos são organizados em diretórios e subdiretórios. O Filestore do Google é baseado em NFS.

Conheça os diferentes tipos de bancos de dados. Os bancos de dados são amplamente divididos em bancos de dados relacionais e NoSQL. Bancos de dados relacionais suportam transações, consistência forte e a linguagem de consulta SQL.

Bancos de dados relacionais têm sido tradicionalmente difíceis de escalar horizontalmente. O Cloud Spanner é um banco de dados relacional global que oferece as vantagens dos bancos de dados relacionais com a escalabilidade anteriormente encontrada apenas em bancos de dados NoSQL.

Bancos de dados NoSQL são projetados para ser escaláveis horizontalmente. Outras características, como consistência forte e suporte para SQL padrão, muitas vezes são sacrificadas para alcançar escalabilidade e respostas de consulta de baixa latência. Bancos de dados NoSQL podem ser lojas de chave-valor como o Cloud Memorystore, bancos de dados de documentos como o Cloud Firestore ou bancos de dados de coluna larga como o Cloud Bigtable.

Entenda as nuvens privadas virtuais. Uma VPC é uma isolamento lógico dos recursos de nuvem de uma organização dentro de uma nuvem pública. No Google Cloud, as VPCs são globais; elas não são restritas a uma única zona ou região. Todo o tráfego entre os serviços do Google Cloud pode ser transmitido pela rede do Google sem a necessidade de enviar tráfego pela Internet pública.

Entenda o balanceamento de carga. O balanceamento de carga é o processo de distribuição de uma carga de trabalho entre um grupo de servidores. Balanceadores de carga podem rotear trabalho com base em regras de nível de rede ou de nível de aplicação. Os平衡adores de carga do Google Cloud podem distribuir cargas de trabalho globalmente.

Entenda as ferramentas de desenvolvedor e gerenciamento. Ferramentas de desenvolvedor suportam fluxos de trabalho comuns em engenharia de software, incluindo o uso de controle de versão para software, construção de contêineres para executar aplicações e serviços, e disponibilização de contêineres para outros desenvolvedores e sistemas de orquestração, como o Kubernetes Engine.

Ferramentas de gerenciamento, como o Cloud Monitoring e o Cloud Logging, são projetadas para fornecer informações de administração de sistemas para desenvolvedores e operadores responsáveis por garantir que as aplicações estejam disponíveis e operando conforme esperado.

Conheça os tipos de serviços especializados oferecidos pelo Google Cloud. O Google Cloud inclui uma lista crescente de serviços especializados para análise de dados, bem como IA e aprendizado de máquina.

Conheça as principais diferenças entre computação on-premises e computação em nuvem pública. A computação on-premises é a computação, armazenamento, rede e serviços relacionados que ocorrem em infraestrutura gerenciada por uma empresa ou organização para seu próprio uso. O hardware pode estar localizado literalmente nas instalações em um prédio da empresa ou em uma instalação de colocation de terceiros. Instalações de colocation fornecem energia, resfriamento e segurança física, mas os clientes da instalação de colocation são responsáveis por toda a configuração e gestão da infraestrutura.

A computação em nuvem pública usa infraestrutura e serviços fornecidos por um provedor de nuvem, como Google, AWS ou Microsoft. O provedor de nuvem mantém

todo o hardware físico e as instalações. Ele fornece uma mistura de serviços, como VMs que são configuradas e mantidas pelos clientes e ofertas sem servidor que permitem aos clientes se concentrar no desenvolvimento de aplicações, enquanto o provedor de nuvem assume mais responsabilidade pela manutenção da infraestrutura de computação subjacente.

Perguntas para Revisão

Você pode encontrar as respostas no Apêndice.

1. Você está planejando implantar uma aplicação SaaS para clientes na América do Norte, Europa e Ásia. Para manter a escalabilidade, você precisará distribuir a carga de trabalho entre servidores em várias regiões. Qual serviço do Google Cloud você usaria para implementar a distribuição de carga de trabalho?
 - A. Cloud DNS
 - B. Cloud Spanner
 - C. Cloud Load Balancing
 - D. Cloud CDN
2. Você decidiu implantar um conjunto de microsserviços usando contêineres. Os microsserviços manterão o estado no contêiner. Você poderia instalar e gerenciar o Docker em instâncias do Compute Engine, mas preferiria que o Google Cloud fornecesse alguns serviços de gerenciamento de contêineres. Quais são dois serviços do Google Cloud que permitem executar contêineres em um serviço gerenciado?
 - A. Ambiente padrão do App Engine e ambiente flexível do App Engine
 - B. Kubernetes Engine e ambiente padrão do App Engine
 - C. Kubernetes Engine e ambiente do Cloud Run
 - D. Ambiente padrão do App Engine e Cloud Functions
3. Por que um desenvolvedor de API gostaria de usar a plataforma API Apigee?
 - A. Para obter os benefícios de roteamento e limitação de taxa
 - B. Serviços de autenticação
 - C. Controle de versão do código
 - D. A e B
 - E. Todas as opções acima
4. Você está implantando uma API na Internet pública e está preocupado que seu serviço esteja sujeito a ataques DDoS. Qual serviço do Google Cloud você deveria considerar para proteger sua API?
 - A. Cloud Armor
 - B. Cloud CDN
 - C. Cloud IAM
 - D. VPCs

5. Você tem uma aplicação que usa uma fila de mensagens Pub/Sub para manter uma lista de tarefas que devem ser processadas por outra aplicação. A aplicação que consome mensagens da fila Pub/Sub remove a mensagem apenas após completar a tarefa. Leva aproximadamente 10 segundos para completar uma tarefa. Não é problema se duas ou mais VMs realizarem a mesma tarefa. Qual é uma configuração custo-efetiva para processar essa carga de trabalho?
 - A. Use VMs preemptivas
 - B. Use VMs padrão
 - C. Use DataProc
 - D. Use Spanner
6. Seu departamento está implantando uma aplicação que possui um banco de dados no back-end. Você está preocupado com a carga de leitura no servidor de banco de dados e deseja ter dados disponíveis na memória para reduzir o tempo de resposta a consultas e reduzir a carga no servidor de banco de dados. Qual serviço do Google Cloud você usaria para manter dados na memória?
 - A. Cloud SQL
 - B. Cloud Memorystore
 - C. Cloud Spanner
 - D. Cloud Firestore
7. O Cloud SDK pode ser usado para configurar e gerenciar recursos em quais dos seguintes serviços?
 - A. Compute Engine
 - B. Cloud Storage
 - C. Firewalls de rede
 - D. Todos os anteriores
8. Qual configuração de servidor é necessária para usar o Cloud Functions?
 - A. Configuração de VM
 - B. Configuração de cluster
 - C. Configuração de Pub/Sub
 - D. Nenhuma
9. Você foi designado para a tarefa de consolidar dados de log gerados por cada instância de uma aplicação. Quais ferramentas de gerenciamento e observabilidade você usaria?
 - A. Cloud Monitoring
 - B. Cloud Trace

- C. Cloud Debugger
 - D. Cloud Logging
10. Quais serviços especializados são mais prováveis de ser usados para construir uma plataforma de armazenamento de dados que requer operações complexas de extração, transformação e carregamento em dados em lote, bem como processamento de dados em streaming?
- A. Plataforma API Apigee
 - B. Análise de dados
 - C. IA e aprendizado de máquina
 - D. Cloud SDK
11. Sua empresa implantou 100.000 sensores de Internet das Coisas (IoT) para coletar dados sobre o estado de equipamentos em várias fábricas. Cada sensor coletará e enviará dados para um armazenamento de dados a cada 5 segundos. Os sensores funcionarão continuamente. Relatórios diários produzirão dados sobre os valores máximos, mínimos e médios para cada métrica coletada em cada sensor. Não há necessidade de suportar transações nesta aplicação. Qual produto de banco de dados você recomendaria?
- A. Cloud Spanner
 - B. Cloud Bigtable
 - C. Cloud SQL MySQL
 - D. Cloud SQL PostgreSQL
12. Você é o desenvolvedor líder em uma aplicação médica que usa smartphones dos pacientes para capturar dados biométricos. O aplicativo é obrigado a coletar dados e armazená-los no smartphone quando os dados não podem ser transmitidos de forma confiável para a aplicação back-end. Você quer minimizar a quantidade de desenvolvimento que você tem que fazer para manter os dados sincronizados entre smartphones e armazenamentos de dados back-end. Qual opção de armazenamento de dados você deveria recomendar?
- A. Cloud Firestore
 - B. Cloud Spanner
 - C. Cloud CDN
 - D. Cloud SQL
13. Um engenheiro de software vem até você para uma recomendação. Eles implementaram um algoritmo de aprendizado de máquina para identificar células cancerígenas em imagens médicas. O algoritmo é computacionalmente intensivo, faz muitos cálculos de ponto flutuante, requer acesso imediato a grandes

quantidades de dados e não pode ser facilmente distribuído por vários servidores. Que tipo de configuração do Compute Engine você recomendaria?

- A. Alta memória, alta CPU
 - B. Alta memória, alta CPU, GPU
 - C. Memória de nível médio, alta CPU
 - D. Alta CPU, GPU
14. Você tem a tarefa de mapear as políticas de autenticação e autorização de suas aplicações locais para os mecanismos de autenticação e autorização do Google Cloud. A documentação do Google Cloud afirma que uma identidade deve ser autenticada para conceder permissões a essa identidade. O que o termo identidade se refere?
- A. ID da VM
 - B. Usuário
 - C. Função
 - D. Conjunto de privilégios
15. Um cliente está desenvolvendo uma aplicação que precisará analisar grandes volumes de informações textuais. O cliente não é especialista em mineração de texto ou trabalho com linguagem. Qual serviço do Google Cloud você recomendaria que eles usassem?
- A. Vertex AI
 - B. Recommendation AI
 - C. Natural Language
 - D. Text-to-Speech
16. Cientistas de dados em sua empresa querem usar uma biblioteca de aprendizado de máquina disponível apenas no Apache Spark. Eles querem minimizar a quantidade de administração e trabalho de DevOps. Como você recomendaria que eles procedessem?
- A. Use Cloud Spark.
 - B. Use Cloud Dataproc.
 - C. Use BigQuery.
 - D. Instale o Apache Spark em um cluster de VMs.
17. Designers de banco de dados em sua empresa estão debatendo a melhor maneira de mover um banco de dados para o Google Cloud. O banco de dados suporta uma aplicação com uma base de usuários global. Os usuários esperam suporte para transações e a capacidade de consultar dados usando ferramentas de consulta comumente usadas. Os designers de banco de dados decidem que qualquer serviço

de banco de dados que escolherem precisará suportar ANSI SQL 2011 e transações globais. Qual serviço de banco de dados você recomendaria?

- A. Cloud SQL
- B. Cloud Spanner
- C. Cloud Firestore
- D. Cloud Bigtable

18. Qual serviço especializado suporta tanto fluxos de trabalho de processamento em lote quanto em streaming?

- A. Cloud Dataflow
- B. BigQuery
- C. Cloud Firestore
- D. AutoML

19. Você tem uma aplicação Python que gostaria de executar em um ambiente escalável com o mínimo de sobrecarga de gerenciamento. Qual produto do Google Cloud você selecionaria?

- A. Ambiente flexível do App Engine
- B. Cloud Engine
- C. Ambiente padrão do App Engine
- D. Kubernetes Engine

20. Um gerente de produto em sua empresa relata que os clientes estão reclamando sobre a confiabilidade de uma de suas aplicações. A aplicação está travando periodicamente, mas os desenvolvedores não encontraram um padrão comum que desencadeie as travadas. Eles estão preocupados que não têm uma boa visão sobre o comportamento da aplicação e querem realizar uma revisão detalhada de todos os dados de travamento. Qual ferramenta de observabilidade você usaria para visualizar informações consolidadas sobre travamentos?

- A. Cloud DataProc
- B. Cloud Monitoring
- C. Cloud Logging
- D. Error Reporting

Capítulo 3

Projetos, Contas de Serviço e Faturamento

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO DO ENGENHEIRO ASSOCIADO À NUVEM DO GOOGLE:

- ✓✓ 1.1 Configuração de projetos e contas na nuvem
- ✓✓ 1.2 Gerenciamento da configuração de faturamento

Antes de nos aprofundarmos em serviços de computação, armazenamento e rede, precisamos discutir como o Google Cloud organiza recursos e vincula o uso desses recursos a um sistema de faturamento. Este capítulo introduz a hierarquia organizacional do Google Cloud, que consiste em organizações, pastas e projetos. Ele também discute contas de serviço, que são maneiras de atribuir funções a recursos de computação para que eles possam realizar funções em seu nome. Finalmente, o capítulo discute brevemente o faturamento.

Como o Google Cloud Organiza Projetos e Contas

Quando você usa o Google Cloud, provavelmente inicia máquinas virtuais ou clusters, talvez crie buckets para armazenar objetos e faça uso de serviços de computação sem servidor, como o Cloud Run e o Cloud Functions. A lista de recursos que você usa pode crescer rapidamente e também pode mudar de maneiras dinâmicas e imprevisíveis à medida que serviços de autoescala respondem à carga de trabalho.

Se você executa uma única aplicação ou alguns serviços para o seu departamento, você pode ser capaz de rastrear todos os recursos visualizando listas de recursos em uso. À medida que o escopo do seu uso do Google Cloud cresce, provavelmente você terá vários departamentos, cada um com seus próprios administradores que precisam de diferentes privilégios. O Google Cloud oferece uma maneira de agrupar recursos e gerenciá-los como uma unidade única. Isso é chamado de hierarquia de recursos. O acesso aos recursos na hierarquia de recursos é controlado por um conjunto de políticas que você pode definir.

Hierarquia de Recursos do Google Cloud

A abstração central para gerenciar recursos do Google Cloud é a hierarquia de recursos. Ela consiste em três níveis:

- Organização
- Pasta
- Projeto

Vamos descrever como esses três componentes se relacionam entre si.

Organização

Uma organização é a raiz da hierarquia de recursos e tipicamente corresponde a uma empresa ou organização. Domínios do Google Workspace e contas do Cloud Identity são mapeados para organizações do Google Cloud. O Google Workspace é o conjunto de produtividade de escritório do Google, que inclui Gmail, Docs, Drive, Calendário e outros serviços. Se sua empresa usa o Google Workspace, você pode criar uma organização na sua hierarquia do Google Cloud. Se sua empresa não usa o Google Workspace, você pode usar o Cloud Identity, a oferta de identidade como serviço (IDaaS) do Google (Figura 3.1).

The screenshot shows the Google Cloud Platform interface. At the top, there's a navigation bar with 'Google Cloud Platform' and 'My First Project'. A search bar says 'Search Products, resources, docs (/)'. On the left, a sidebar titled 'IAM & Admin' has several options: 'IAM', 'Identity & Organization' (which is selected and highlighted in blue), 'Policy Troubleshooter', 'Policy Analyzer', 'Organization Policies', 'Service Accounts', 'Workload Identity Federat...', and 'Labels'. The main content area is titled 'Identity' and contains a section titled 'Set up Google Cloud for your organization'. It includes a brief description of the setup checklist and a bulleted list of actions: 'Manage user accounts and groups for employees', 'Create organizational structure and centrally control all of your organization's projects and resources', and 'Configure security guardrails'. Below this is a 'GO TO THE CHECKLIST' button.

Uma única identidade na nuvem está associada a, no máximo, uma organização. Identidades na nuvem têm super administradores, e esses super administradores atribuem o papel de Administrador da Organização de Identidade e Gerenciamento de Acesso (IAM) aos usuários que gerenciam a organização. Além disso, o Google Cloud automaticamente concederá papéis IAM de Criador de Projeto e Criador de Conta de Cobrança a todos os usuários no domínio. Isso permite que qualquer usuário crie projetos e habilite a cobrança pelo custo dos recursos usados.

Os usuários com o papel IAM de Administrador da Organização são responsáveis pelo seguinte:

- Definir a estrutura da hierarquia de recursos
- Definir políticas de identidade e gerenciamento de acesso sobre a hierarquia de recursos
- Delegar outros papéis de gerenciamento a outros usuários

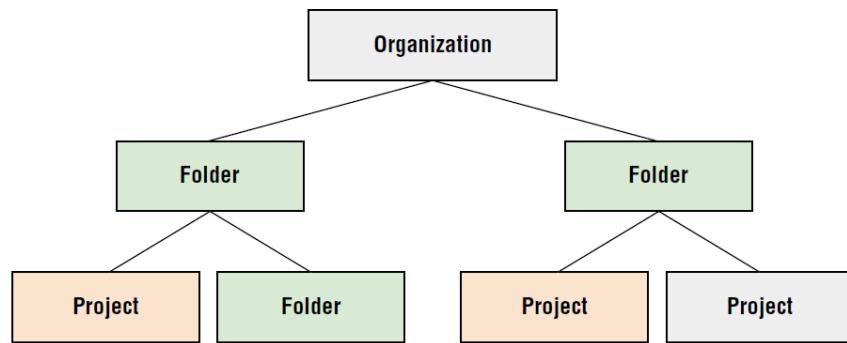
Quando um membro de uma organização do Google Workspace/conta do Cloud Identity cria uma conta de cobrança ou projeto, o Google Cloud automaticamente cria um recurso de organização.

Todos os projetos e contas de faturamento serão filhos do recurso de organização. Além disso, quando a organização é criada, todos os usuários dessa organização recebem as funções de Criador de Projeto e Criador de Conta de Faturamento. A partir desse ponto, os usuários do Google Workspace terão acesso aos recursos do Google Cloud.

Pasta

Pastas são os blocos de construção de hierarquias organizacionais multilayer. Organizações contêm pastas. Pastas podem conter outras pastas ou projetos. No entanto, as pastas são opcionais e não precisam ser usadas. Uma única pasta pode conter tanto pastas quanto projetos (veja a Figura 3.2). A organização de pastas é geralmente construída em torno dos tipos de serviços fornecidos pelos recursos nos projetos contidos e as políticas que governam pastas e projetos.

FIGURE 3.2 Generic organization folder project



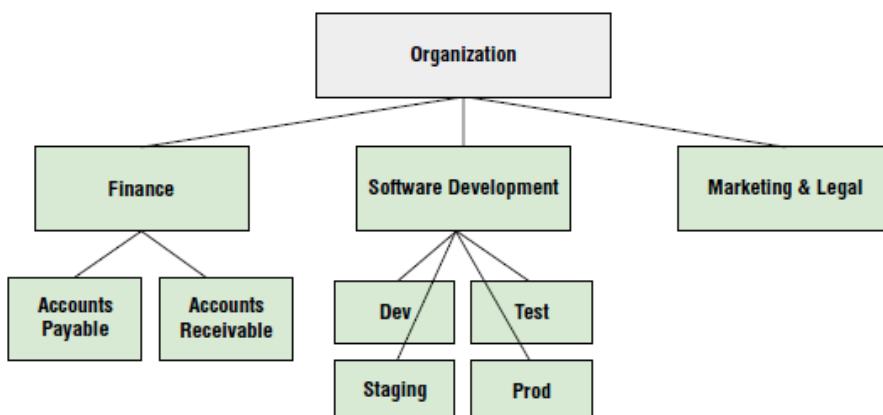
Considere um exemplo de hierarquia de recursos. Uma organização tem quatro departamentos: financeiro, marketing, desenvolvimento de software e jurídico. O departamento financeiro precisa manter seus recursos de contas a receber e a pagar separados, então o administrador cria duas pastas dentro da pasta Financeira: Contas a Receber e Contas a Pagar. O desenvolvimento de software utiliza vários ambientes, incluindo Desenvolvimento, Teste, Staging e Produção. O acesso a cada um dos ambientes é controlado por políticas específicas àquele ambiente, então faz sentido organizar cada ambiente em sua própria pasta. Marketing e jurídico podem ter todos os seus recursos compartilhados entre os membros do departamento, então uma única pasta é suficiente para ambos os departamentos. A Figura 3.3 mostra a hierarquia organizacional para esta organização.

Agora que definimos uma organização e configuramos pastas que correspondem aos nossos departamentos e como diferentes grupos de recursos serão acessados, podemos criar projetos.

Projeto

Projetos são, de certa forma, a parte mais importante da hierarquia. É nos projetos que criamos recursos, usamos serviços do Google Cloud, gerenciamos permissões e opções de faturamento.

FIGURE 3.3 Example organization folder project



O primeiro passo para trabalhar com um projeto é criar um. Qualquer pessoa com a permissão resourcemanager.projects.create do IAM pode criar um projeto. Por padrão, quando uma organização é criada, cada usuário no domínio recebe essa permissão. Sua organização terá uma cota de projetos que pode criar. A cota pode variar entre organizações. O Google toma decisões sobre cotas de projetos com base no uso típico, no histórico de uso do cliente e em outros fatores. Se você atingir seu limite de projetos e tentar criar outro, será solicitado a solicitar um aumento na cota. Você terá que fornecer informações como o número de projetos adicionais de que precisa e para que serão usados.

Depois de criar sua hierarquia de recursos, você pode definir políticas que a governam.

Políticas da Organização

O Google Cloud fornece um Serviço de Política da Organização. Este serviço controla o acesso aos recursos de uma organização. O Serviço de Política da Organização complementa o serviço de IAM. O IAM permite atribuir permissões para que usuários ou funções possam realizar operações específicas na nuvem. O Serviço de Política da Organização permite especificar limites nas maneiras como os recursos podem ser usados. Uma maneira de pensar na diferença é que o IAM especifica quem pode fazer as coisas, e o Serviço de Política da Organização especifica o que pode ser feito com os recursos.

As políticas da organização são definidas em termos de restrições a um recurso.

Restrições nos Recursos

Restrições são limitações nos serviços. O Google Cloud possui restrições de lista e restrições booleanas.

As restrições de lista são listas de valores que são permitidos ou negados para um recurso. Os seguintes são alguns tipos de restrições de lista:

- Permitir um conjunto específico de valores.
- Negar um conjunto específico de valores.
- Negar um valor e todos os seus valores filhos.
- Permitir todos os valores permitidos.
- Negar todos os valores.

Restrições booleanas avaliam como verdadeiro ou falso e determinam se a restrição é aplicada ou não. Por exemplo, se você quiser negar o acesso às portas seriais em VMs, você pode definir constraints/compute.disableSerialPortAccess como TRUE.

Veja a documentação de restrições de política da organização em <https://cloud.google.com/resource-manager/docs/organization-policy/org-policy-constraints> para mais detalhes.

Avaliação de Política

Organizações podem ter políticas permanentes para proteger dados e recursos na nuvem. Por exemplo, pode haver regras que ditam quem na organização pode habilitar uma API de serviço ou criar uma conta de serviço. Seu departamento de InfoSec pode exigir que todas as VMs desativem o acesso à porta serial. Você poderia implementar controles em cada VM individualmente, mas isso é ineficiente e propenso a erros. Uma abordagem melhor é definir uma política que restrinja o que pode ser feito e anexar essa política a um objeto na hierarquia de recursos.

Por exemplo, já que o InfoSec quer que todas as VMs desativem o acesso à porta serial, você poderia especificar uma política que restrinja o acesso à porta serial e, em seguida, anexá-la à organização. Todas as pastas e projetos abaixo da organização herdarão essa política. Como as políticas são herdadas e não podem ser desabilitadas ou substituídas por objetos mais baixos na hierarquia, esta é uma forma eficaz de aplicar uma política em todos os recursos organizacionais. No entanto, há uma maneira de desabilitar a herança dos pais definindo o parâmetro `inheritFromParent` como falso.

As políticas são gerenciadas através do formulário de Políticas da Organização no console IAM & Admin. A Figura 3.4 mostra um conjunto de políticas de exemplo.

Múltiplas políticas podem estar em vigor para uma pasta ou projeto. Por exemplo, se a organização tivesse uma política sobre o acesso à porta serial e uma pasta contendo um projeto tivesse uma política limitando quem pode criar contas de serviço, então o projeto herdaria ambas as políticas e ambas restringiriam o que poderia ser feito com recursos naquele projeto.

Gerenciamento de Projetos

Uma das primeiras tarefas que você realizará ao iniciar uma nova iniciativa na nuvem é configurar um projeto. Isso pode ser feito com o Google Cloud Console. Supondo que você tenha criado uma conta com o Google Cloud, navegue até o Google Cloud Console em <https://console.cloud.google.com> e faça login. Você verá a página inicial, que se parece com a Figura 3.5.

FIGURE 3.4 Organizational policies are managed in the IAM & Admin console.

The screenshot shows the Google Cloud Platform IAM & Admin console. The left sidebar has a navigation menu with 'Organization Policies' selected. The main area displays a table of organization policies for the project 'My First Project'. The table columns are 'Name', 'ID', and 'Inheritance'. The policies listed are:

Name	ID	Inheritance
Allow extending lifetime of OAuth 2.0 access tokens to up to 12 hours	constraints/iam.allowServiceAccountCredentialLifetimeExtension	Inherited
Allowed AWS accounts that can be configured for workload identity federation in Cloud IAM	constraints/iam.workloadIdentityPoolAwsAccounts	Inherited
Allowed Binary Authorization Policies (Cloud Run)	constraints/run.allowedBinaryAuthorizationPolicies	Inherited
Allowed Destinations for Exporting Resources	constraints/resourcemanager.allowedExportDestinations	Inherited
Allowed external identity Providers for workloads in Cloud IAM	constraints/iam.workloadIdentityPoolProviders	Inherited
Allowed ingress settings (Cloud Functions)	constraints/cloudfunctions.allowedIngressSettings	Inherited

FIGURE 3.5 Home page console

The screenshot shows the Google Cloud Platform Home page. The top navigation bar includes 'DASHBOARD', 'ACTIVITY', 'RECOMMENDATIONS', and 'CUSTOMIZE'. The main content area is divided into several sections:

- Project info:** Shows the project name 'My First Project', project number '388947348090', and project ID 'scenic-energy-335022'. It also has a link to 'Go to project settings'.
- RPC APIs:** A chart titled 'Requests (requests/sec)' showing data for the selected time frame. The chart has four data series: 1.0, 0.8, 0.6, and 0.4. Below the chart is a link to 'Go to APIs overview'.
- Google Cloud Platform status:** Shows 'All services normal' and a link to 'Go to Cloud status dashboard'.
- Billing:** Shows estimated charges of '\$0.00' for the period Jan 1 – 29, 2022. It includes links to 'Take a tour of billing' and 'View detailed charges'.
- Monitoring:** Includes links to 'Create my dashboard', 'Set up alerting policies', 'Create uptime checks', and 'View all dashboards'.
- Resources:** A list of services including BigQuery, SQL, Compute Engine, Storage, Cloud Functions, and App Engine.

No menu de navegação no canto superior esquerdo, selecione IAM & Admin e, em seguida, selecione Gerenciar Recursos (veja a Figura 3.6 e a Figura 3.7).

FIGURE 3.6 Navigation menu

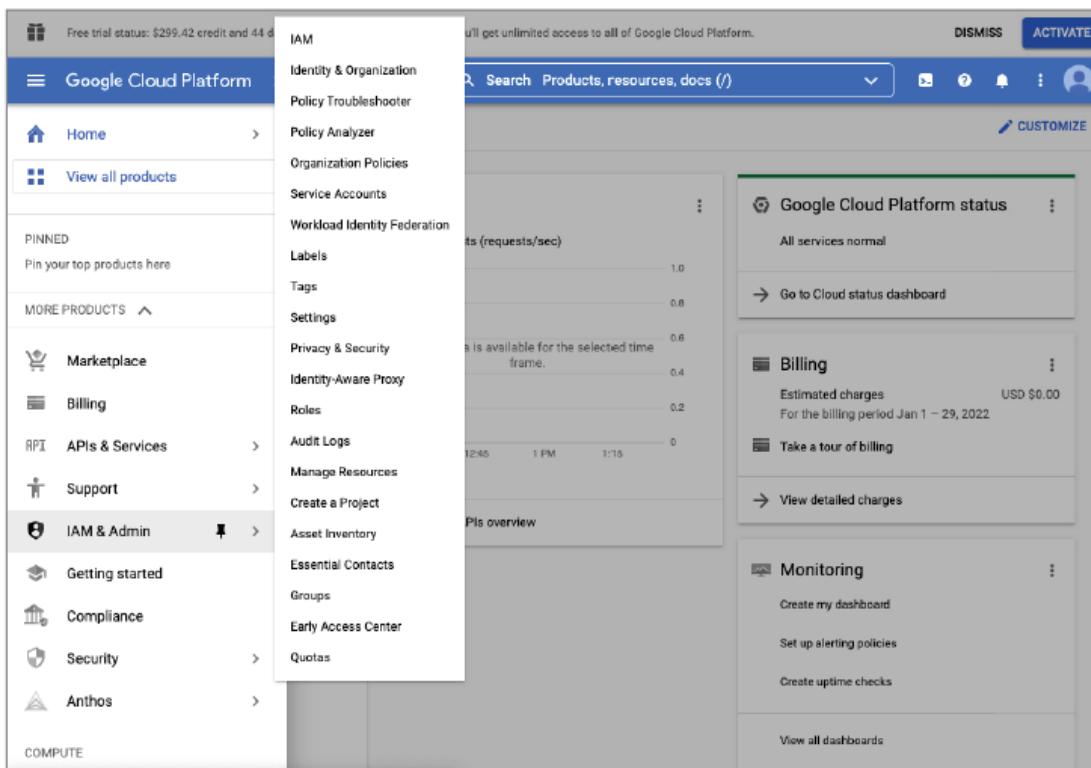
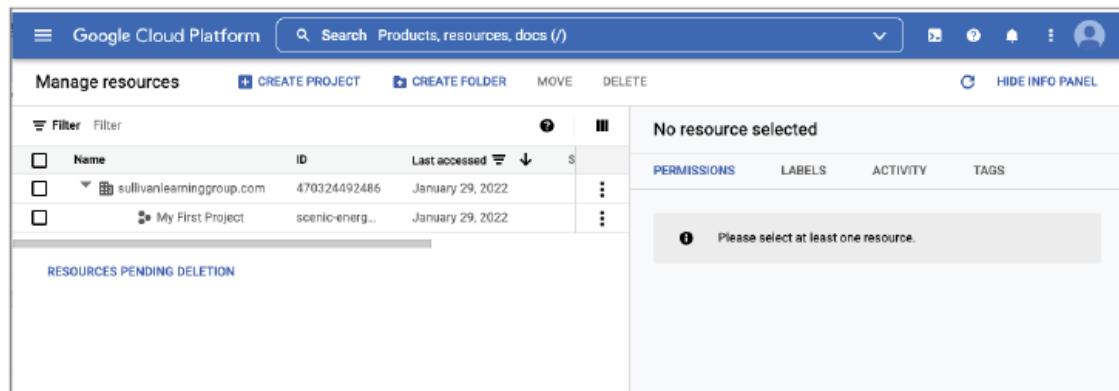


FIGURE 3.7 Managing resources



A partir daí, você pode clicar em Criar Projeto, o que exibirá a caixa de diálogo Novo Projeto. Aqui, você pode inserir o nome de um projeto e selecionar uma organização (Figura 3.8 e Figura 3.9).

FIGURE 3.8 Click Create Project.

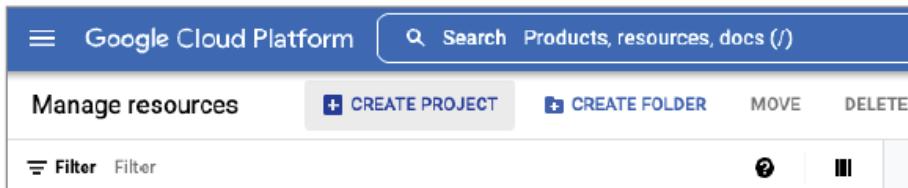


FIGURE 3.9 New Project dialog box

A screenshot of the 'New Project' dialog box. It includes fields for 'Project name *' (containing 'My Project 3502'), 'Organization *' (containing 'sullivanlearninggroup.com'), and 'Location *' (containing 'sullivanlearninggroup.com'). There are 'CREATE' and 'CANCEL' buttons at the bottom. A note above the organization field states: 'Project ID: silken-butress-339721. It cannot be changed later.' and 'Select an organization to attach it to a project. This selection can't be changed later.'

Observe que, ao criar um projeto, sua cota restante de projetos é exibida. Se precisar de projetos adicionais, clique no link Gerenciar cotas para solicitar um aumento em sua cota.

Funções e Identidades

Além de gerenciar recursos, como engenheiro de nuvem você terá que gerenciar o acesso a esses recursos. Isso é feito com o uso de papéis e identidades.

Um papel é uma coleção de permissões. Os papéis são concedidos aos usuários vinculando um usuário a um papel. Quando falamos de identidades, nos referimos ao objeto que usamos para representar um usuário humano ou conta de serviço no Google Cloud. Por exemplo, Alice é uma engenheira de software desenvolvendo aplicações na nuvem (o usuário humano), e ela tem uma identidade com o nome alice@example.com. Papéis são atribuídos a alice@example.com dentro do Google Cloud para que Alice possa criar, modificar, deletar e usar recursos no Google Cloud.

Existem três tipos de papéis no Google Cloud:

- Papéis Básicos
- Papéis Predefinidos

- Papéis Personalizados

Papéis básicos, anteriormente conhecidos como papéis primitivos, incluem Dono, Editor e Visualizador. Estes fornecem privilégios amplos que podem ser aplicados à maioria dos recursos. É uma melhor prática usar papéis predefinidos em vez de papéis básicos quando possível. Papéis básicos concedem amplas faixas de permissões que podem não ser sempre necessárias por um usuário. Ao usar papéis predefinidos, você pode conceder apenas as permissões que um usuário precisa para realizar sua função. Esta prática de apenas atribuir as permissões necessárias e nada mais é conhecida como o princípio do menor privilégio. É uma das práticas fundamentais na segurança da informação.

Papéis predefinidos fornecem acesso granular aos recursos no Google Cloud, e são específicos para produtos do Google Cloud e gerenciados e atualizados pelo Google. Por exemplo, os papéis do App Engine incluem o seguinte:

- appengine.appAdmin, que concede às identidades a habilidade de ler, escrever e modificar todas as configurações da aplicação no App Engine
- appengine.ServiceAdmin, que concede acesso somente leitura às configurações da aplicação e acesso de nível de escrita às configurações de módulo e de versão no App Engine
- appengine.appViewer, que concede acesso somente leitura às aplicações no App Engine

Papéis personalizados permitem que os administradores da nuvem criem e administrem seus próprios papéis. Papéis personalizados são montados usando permissões definidas no IAM. Enquanto você pode usar a maioria das permissões em um papel personalizado, algumas, como iam.ServiceAccounts.getAccessToken, não estão disponíveis em papéis personalizados.

Concessão de funções a identidades

Concedendo Papéis a Identidades Uma vez que você tenha determinado quais papéis deseja fornecer aos usuários, você pode atribuir papéis aos usuários através do console IAM. É importante saber que permissões não podem ser atribuídas diretamente aos usuários — elas só podem ser atribuídas a papéis. Os papéis são então atribuídos aos usuários.

Do console IAM, você pode selecionar um projeto que exibirá uma interface de permissão, como na Figura 3.11. A partir daí, selecione a opção Adicionar para exibir outra caixa de diálogo que solicita nomes de usuários e papéis (veja Figura 3.12).

FIGURE 3.10 A sample list of roles in Google Cloud

The screenshot shows the Google Cloud Platform interface for managing IAM roles. The left sidebar under 'IAM & Admin' has 'Roles' selected. The main area displays a table of roles for the project 'My First Project'. The table columns are Type, Title, Used in, and Status. The roles listed are: AAM Admin, AAM Conversational Architect, AAM Dialog Designer, AAM Lead Dialog Designer, AAM Viewer, Access Approval Approver, Access Approval Config Editor, Access Approval Viewer, Access Context Manager Admin, Access Context Manager Editor, Access Context Manager Reader, Access Transparency Admin, Actions Admin, Actions Viewer, and Activity Analysis Viewer. All roles are currently enabled.

Type	Title	Used in	Status
Cloud IAM	AAM Admin	Dialogflow	Enabled
Cloud IAM	AAM Conversational Architect	Dialogflow	Enabled
Cloud IAM	AAM Dialog Designer	Dialogflow	Enabled
Cloud IAM	AAM Lead Dialog Designer	Dialogflow	Enabled
Cloud IAM	AAM Viewer	Dialogflow	Enabled
Cloud IAM	Access Approval Approver	Access Approval	Enabled
Cloud IAM	Access Approval Config Editor	Access Approval	Enabled
Cloud IAM	Access Approval Viewer	Access Approval	Enabled
Cloud IAM	Access Context Manager Admin	Access Context Manager	Enabled
Cloud IAM	Access Context Manager Editor	Access Context Manager	Enabled
Cloud IAM	Access Context Manager Reader	Access Context Manager	Enabled
Cloud IAM	Access Transparency Admin	Organization Policy	Enabled
Cloud IAM	Actions Admin	Actions	Enabled
Cloud IAM	Actions Viewer	Actions	Enabled
Cloud IAM	Activity Analysis Viewer	Other	Enabled

FIGURE 3.11 IAM permissions

The screenshot shows the Google Cloud Platform interface for managing IAM permissions. The left sidebar under 'IAM & Admin' has 'Permissions' selected. The main area displays a table of permissions for the project 'My First Project'. The table columns are Type, Principal, Name, and Role. One permission is listed: 'Compute Engine default service account' with the role 'Editor'. There is also a note indicating that this permission affects all resources in the project.

Type	Principal ↑	Name	Role
Cloud IAM	388947348090-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor

FIGURE 3.12 Adding a user

The screenshot shows a dialog box titled 'Add principals to "My First Project"'. Below it, a sub-section titled 'Add principals and roles for "My First Project" resource' contains instructions: 'Enter one or more principals below. Then select a role for these principals to grant them access to your resources. Multiple roles allowed.' A link 'Learn more' is provided. A text input field labeled 'New principals' is highlighted with a blue border. To its right is a question mark icon. Below this is a dropdown menu labeled 'Select a role' with a downward arrow. To the right of the dropdown is a 'Condition' section with a 'Add condition' link and a trash can icon. At the bottom left are 'SAVE' and 'CANCEL' buttons.

Contas de Serviço

Identidades são geralmente associadas a usuários individuais. Às vezes, é útil ter aplicações ou VMs agindo em nome de um usuário ou realizando operações que o usuário não tem permissão para executar.

Por exemplo, você pode ter uma aplicação que precisa acessar um banco de dados, mas não deseja permitir que os usuários da aplicação acessem o banco de dados diretamente. Em vez disso, todas as solicitações de usuário para o banco de dados devem passar pela aplicação. Você pode criar uma conta de serviço que tenha acesso ao banco de dados. Você pode então atribuir essa conta de serviço à aplicação, de modo que a aplicação possa executar consultas em nome dos usuários sem ter que conceder acesso ao banco de dados a esses usuários.

Contas de serviço são um pouco incomuns no sentido de que às vezes as tratamos como recursos e às vezes como identidades. Quando atribuímos um papel a uma conta de serviço, estamos tratando-a como uma identidade. Quando damos permissão aos usuários para acessar uma conta de serviço, estamos tratando-a como um recurso.

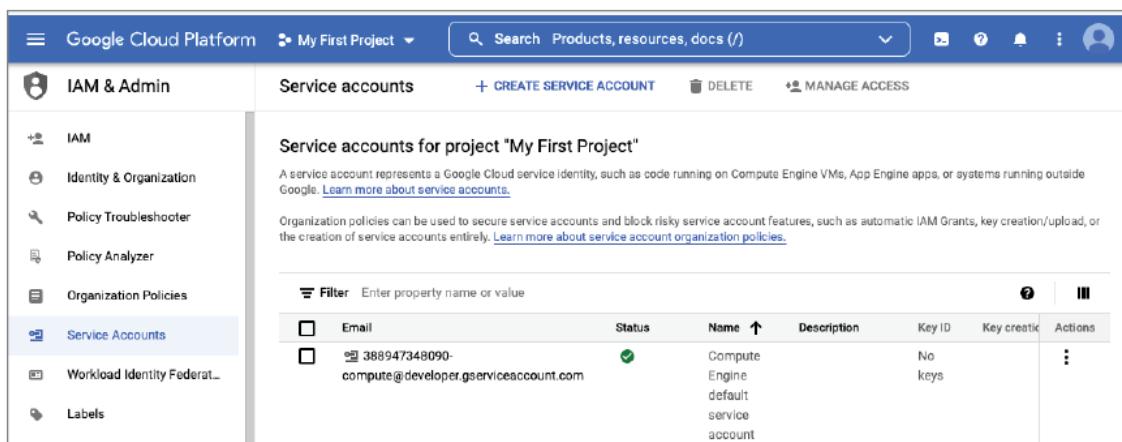
Existem dois tipos de contas de serviço: contas de serviço gerenciadas pelo usuário e contas de serviço gerenciadas pelo Google. Os usuários podem criar até 100 contas de serviço por projeto. Quando você cria um projeto que tem a API do Compute Engine habilitada, uma conta de serviço do Compute Engine é criada automaticamente. Da mesma forma, se você tem uma aplicação App Engine em seu projeto, o Google Cloud criará automaticamente uma conta de serviço do App Engine. Tanto as contas de serviço do Compute Engine quanto do App Engine recebem papéis de editor nos projetos nos

quais são criadas. Você também pode criar contas de serviço personalizadas em seus projetos.

O Google também pode criar contas de serviço que ele gerencia. Essas contas são usadas com vários serviços do Google Cloud.

Contas de serviço podem ser gerenciadas como um grupo de contas no nível do projeto ou no nível individual da conta de serviço. Por exemplo, se você conceder iam.serviceAccountUser a um usuário para um projeto específico, então esse usuário pode gerenciar todas as contas de serviço no projeto. Se preferir limitar os usuários a gerenciar apenas contas de serviço específicas, você poderia conceder iam.serviceAccountUser para uma conta de serviço específica.

Contas de serviço são criadas automaticamente quando os recursos são criados. Por exemplo, uma conta de serviço será criada para uma VM quando a VM é criada. Pode haver situações nas quais você gostaria de criar uma conta de serviço para uma de suas aplicações. Nesse caso, você pode navegar até o console IAM & Admin e selecionar Contas de Serviço. A partir daí, você pode clicar em Criar Conta de Serviço no topo, conforme mostrado na Figura 3.13.



The screenshot shows the Google Cloud Platform interface for managing service accounts. The left sidebar is titled 'IAM & Admin' and includes options like IAM, Identity & Organization, Policy Troubleshooter, Policy Analyzer, Organization Policies, Service Accounts (which is selected), Workload Identity Federation, and Labels. The main content area is titled 'Service accounts for project "My First Project"'. It contains a brief description: 'A service account represents a Google Cloud service identity, such as code running on Compute Engine VMs, App Engine apps, or systems running outside Google. [Learn more about service accounts](#).'. Below this, another note says: 'Organization policies can be used to secure service accounts and block risky service account features, such as automatic IAM Grants, key creation/upload, or the creation of service accounts entirely. [Learn more about service account organization policies](#)'. A table lists existing service accounts:

Email	Status	Name	Description	Key ID	Key creation	Actions
388947348090-compute@developer.gserviceaccount.com	✓	Compute Engine default service account	No keys			⋮

Isso traz um formulário que solicita as informações necessárias para criar uma conta de serviço.

Faturamento

Usar recursos como VMs, armazenamento de objetos e serviços especializados geralmente incorre em cobranças. A API de Faturamento do Google Cloud fornece uma maneira de gerenciar como você paga pelos recursos usados.

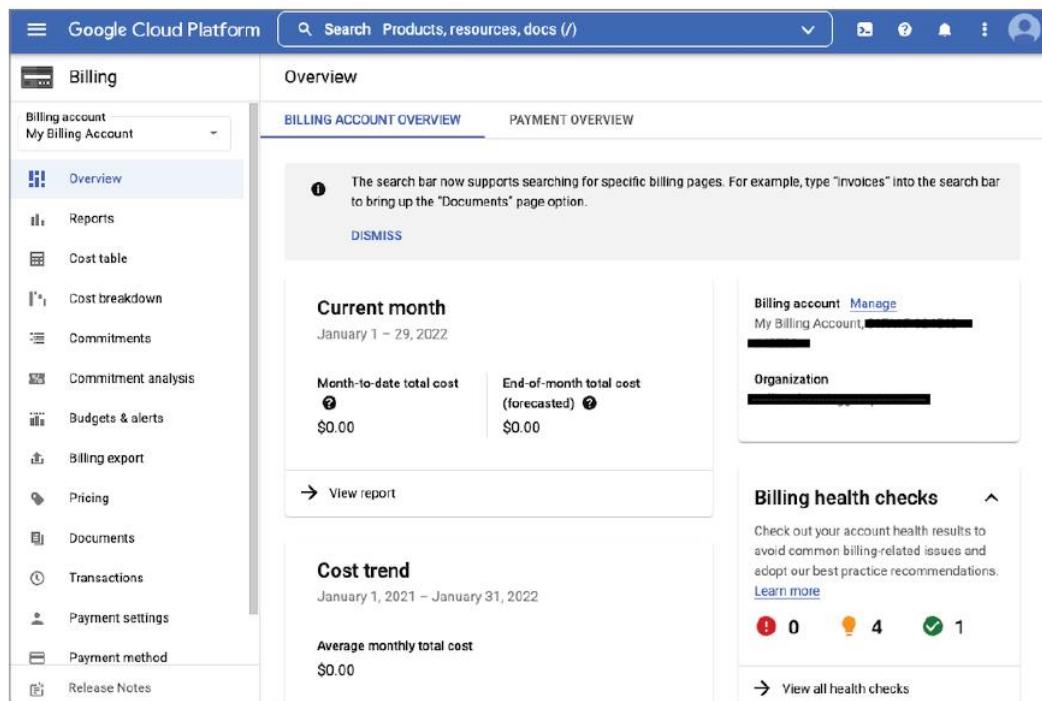
Contas de Faturamento

Contas de faturamento armazenam informações sobre como pagar as cobranças pelos recursos usados. Uma conta de faturamento está associada a um ou mais projetos. Todos os projetos devem ter uma conta de faturamento, a menos que usem apenas serviços gratuitos.

Contas de faturamento podem seguir uma estrutura semelhante à hierarquia de recursos. Se você estiver trabalhando com uma empresa pequena, pode ter apenas uma única conta de faturamento. Nesse caso, todos os custos de recursos são cobrados dessa única conta. Se sua empresa for semelhante ao exemplo mencionado anteriormente no capítulo, com departamentos financeiro, de marketing, jurídico e de desenvolvimento de software, então você pode querer ter várias contas de faturamento. Você poderia ter uma conta de faturamento para cada departamento, mas isso pode não ser necessário. Se os serviços de nuvem de finanças, marketing e jurídico forem pagos pela mesma parte do orçamento da sua empresa, eles poderiam usar uma única conta de faturamento. Se os serviços de desenvolvimento de software forem pagos de uma parte diferente do orçamento da sua empresa, então eles poderiam usar uma conta de faturamento diferente.

A partir do console principal do Google Cloud, você pode navegar até o console de Faturamento (veja a Figura 3.14), que lista as contas de faturamento existentes.

FIGURE 3.14 The main Billing form listing existing billing accounts

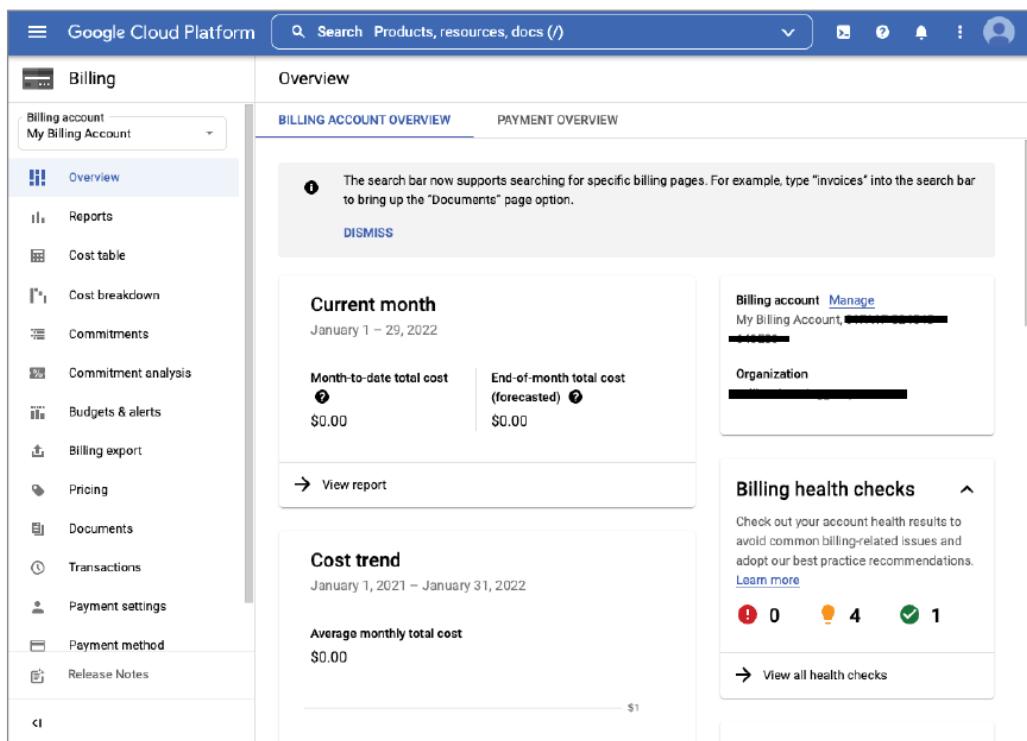


A partir daqui, você pode criar uma nova conta de faturamento, conforme mostrado na Figura 3.15.

Na página de visão geral do Faturamento, você pode visualizar e modificar projetos vinculados às contas de faturamento.

Existem dois tipos de contas de faturamento: autoatendimento e faturadas. As contas de autoatendimento são pagas por cartão de crédito ou débito direto de uma conta bancária. Os custos são cobrados automaticamente. O outro tipo é uma conta de faturamento faturada, na qual faturas são enviadas aos clientes. Esse tipo de conta é comumente usado por empresas e outros grandes clientes.

FIGURE 3.15 The form to create a new billing account



Vários papéis estão associados ao faturamento. É importante conhecê-los para o exame. Os papéis de faturamento são os seguintes:

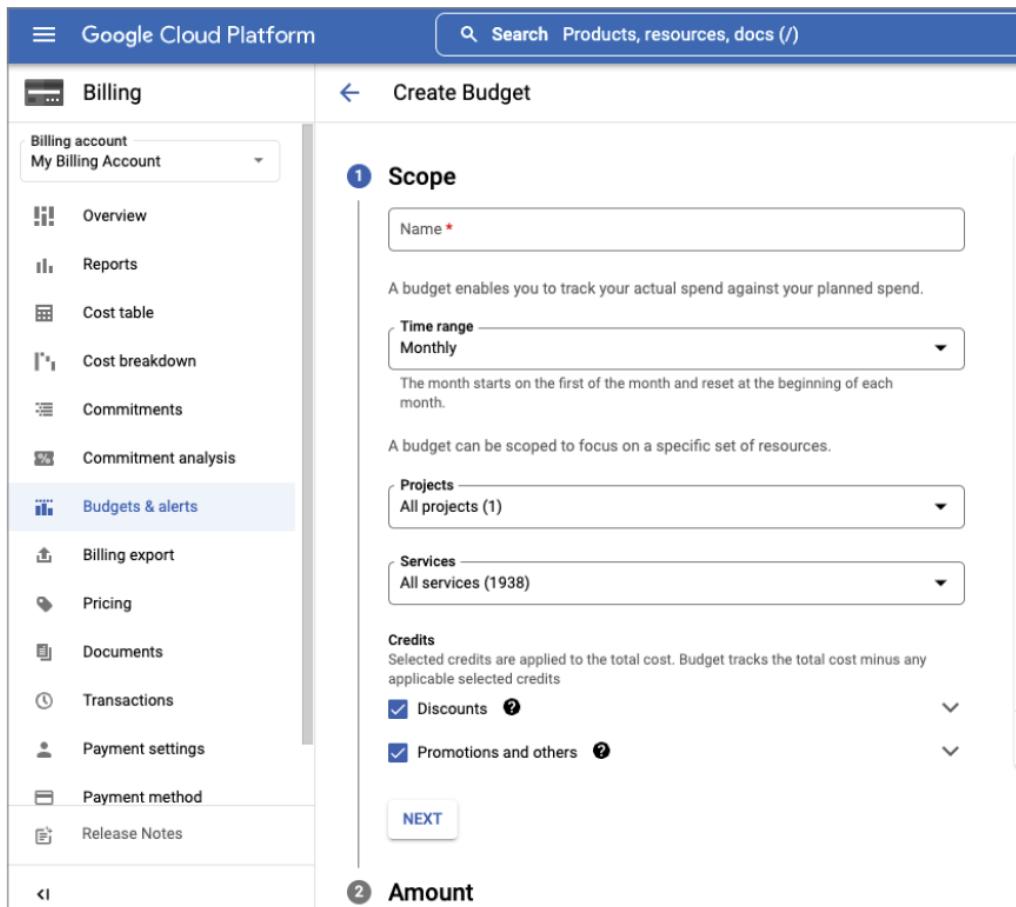
- Criador de Conta de Faturamento, que pode criar novas contas de faturamento de autoatendimento
- Administrador de Conta de Faturamento, que gerencia contas de faturamento mas não pode criá-las
- Usuário de Conta de Faturamento, que permite a um usuário vincular projetos a contas de faturamento
- Visualizador de Conta de Faturamento, que permite a um usuário visualizar custos de conta de faturamento e transações

Poucos usuários provavelmente terão o papel de Criador de Conta de Faturamento, e aqueles que têm provavelmente desempenham um papel financeiro na organização. Administradores de nuvem podem ter o papel de Administrador de Conta de Faturamento para gerenciar as contas. Qualquer usuário que possa criar um projeto deve ter o papel de Usuário de Conta de Faturamento para que novos projetos possam ser vinculados à conta de faturamento apropriada. Visualizador de Conta de Faturamento é útil para alguns, como um auditor que precisa ser capaz de ler informações da conta de faturamento mas não alterá-las.

Orçamentos e Alertas de Faturamento

O serviço de Faturamento do Google Cloud inclui uma opção para definir um orçamento e configurar alertas de faturamento. Você pode navegar até o formulário de orçamento a partir do menu principal do console, selecionar Faturamento e, em seguida, selecionar Orçamentos & Alertas (veja a Figura 3.16).

FIGURE 3.16 The budget form enables you to have notices sent to you when certain percentages of your budget have been spent in a particular month.



No formulário de orçamento, você pode nomear seu orçamento e especificar uma conta de faturamento para monitorar. Note que um orçamento está associado a uma conta de faturamento, não a um projeto. Um ou mais projetos podem ser vinculados a uma conta de faturamento, então o orçamento e os alertas que você especificar devem ser baseados no que você espera gastar para todos os projetos vinculados à conta de faturamento.

FIGURA 3.16 O formulário de orçamento permite que você receba notificações quando certas porcentagens do seu orçamento foram gastas em um determinado mês.

Você pode especificar um valor específico ou especificar que seu orçamento é o montante gasto no mês anterior.

Com um orçamento, você pode definir múltiplas porcentagens de alerta. Por padrão, três porcentagens são definidas: 50%, 90% e 100%. Você pode alterá-las para as porcentagens que funcionam melhor para você. Se desejar mais de três alertas, você pode clicar em Adicionar Item na seção Definir Alertas de Orçamento para adicionar limites de alerta adicionais.

Quando essa porcentagem do orçamento foi gasta, notificará os administradores de faturamento e os usuários da conta de faturamento por e-mail. Se você gostaria de responder aos alertas programaticamente, você pode fazer com que as notificações sejam

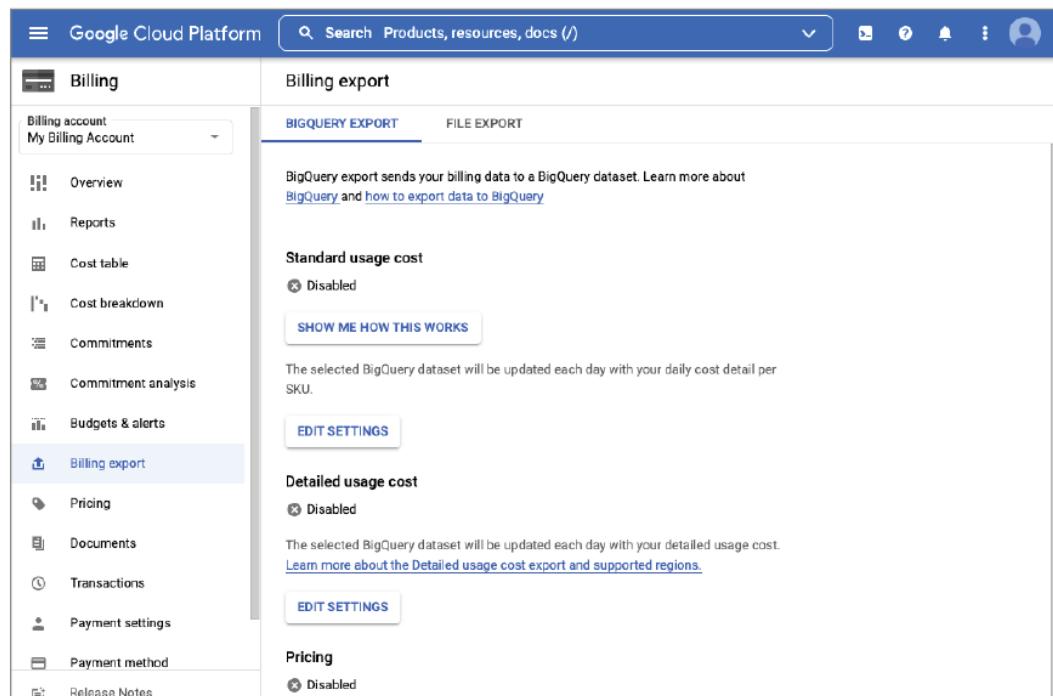
enviadas para um tópico do Pub/Sub marcando a caixa apropriada nas seções Gerenciar Notificação.

Exportando Dados de Faturamento

Você pode exportar dados de faturamento para análise posterior ou por motivos de conformidade. Os dados de faturamento podem ser exportados para o BigQuery.

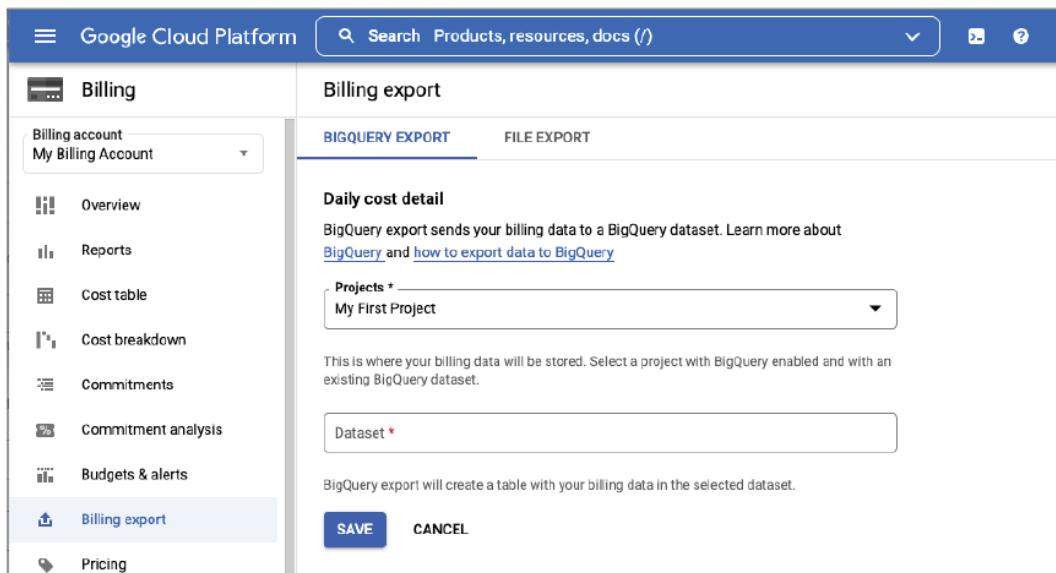
Para exportar dados de faturamento para o BigQuery, navegue até a seção de Faturamento do console e selecione Exportação de Faturamento no menu. No formulário que aparece, selecione a conta de faturamento que você gostaria de exportar (veja a Figura 3.17).

FIGURE 3.17 Billing export form



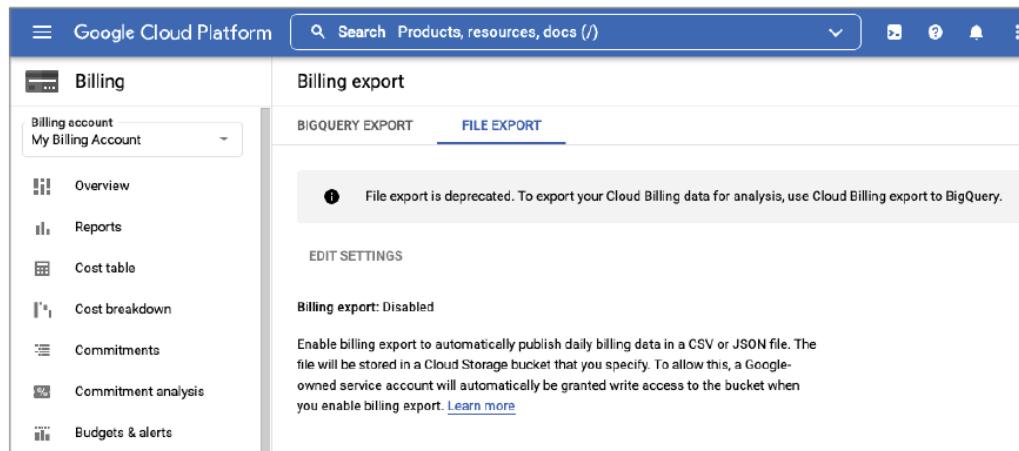
Para o BigQuery, clique em Editar Configuração. Selecione os projetos que deseja incluir. Você precisará criar um conjunto de dados no BigQuery para armazenar os dados. Clique em Ir Para o BigQuery para abrir um formulário do BigQuery. Isso criará um conjunto de dados de exportação de faturamento, que será usado para armazenar os dados exportados. (Veja a Figura 3.18.) Para informações adicionais sobre o uso do BigQuery, veja o Capítulo 12, "Implantando Armazenamento no Google Cloud."

FIGURE 3.18 Exporting to BigQuery



Alternativamente, no passado, você poderia exportar dados de faturamento para um arquivo armazenado no Cloud Storage, mas isso não é mais suportado. Uma opção de Exportação de Arquivo está disponível, mas ela não funciona mais, como mostrado na Figura 3.19. Até o momento em que você ler isto, a opção de Exportação de Arquivo pode ter sido removida.

FIGURE 3.19 Exporting billing data to a file is now deprecated.



Ao exportar para um arquivo, você precisará especificar um nome de bucket e um prefixo de relatório. Você tem a opção de escolher entre o formato de arquivo CSV ou JSON. Pode haver questões no exame sobre as opções de formato de arquivo disponíveis, portanto, lembre-se dessas duas opções.

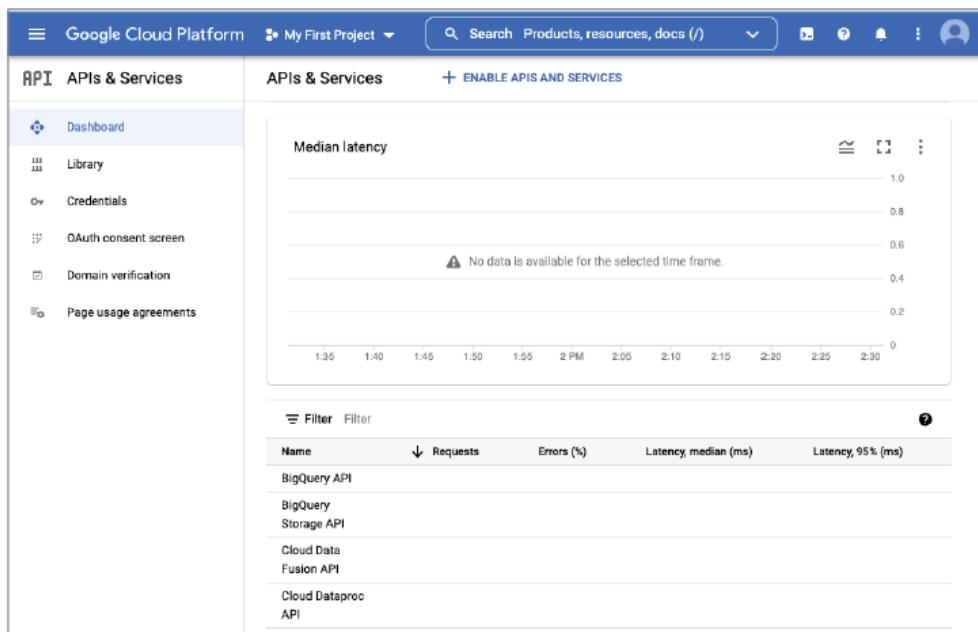
Habilitando APIs

O Google Cloud usa APIs para tornar os serviços acessíveis programaticamente. Por exemplo, quando você usa um formulário para criar uma VM ou um bucket do Cloud Storage, por trás dos bastidores, funções da API são executadas para criar a VM ou o

bucket. Todos os serviços do Google Cloud têm APIs associadas a eles. No entanto, a maioria não está habilitada por padrão em um projeto.

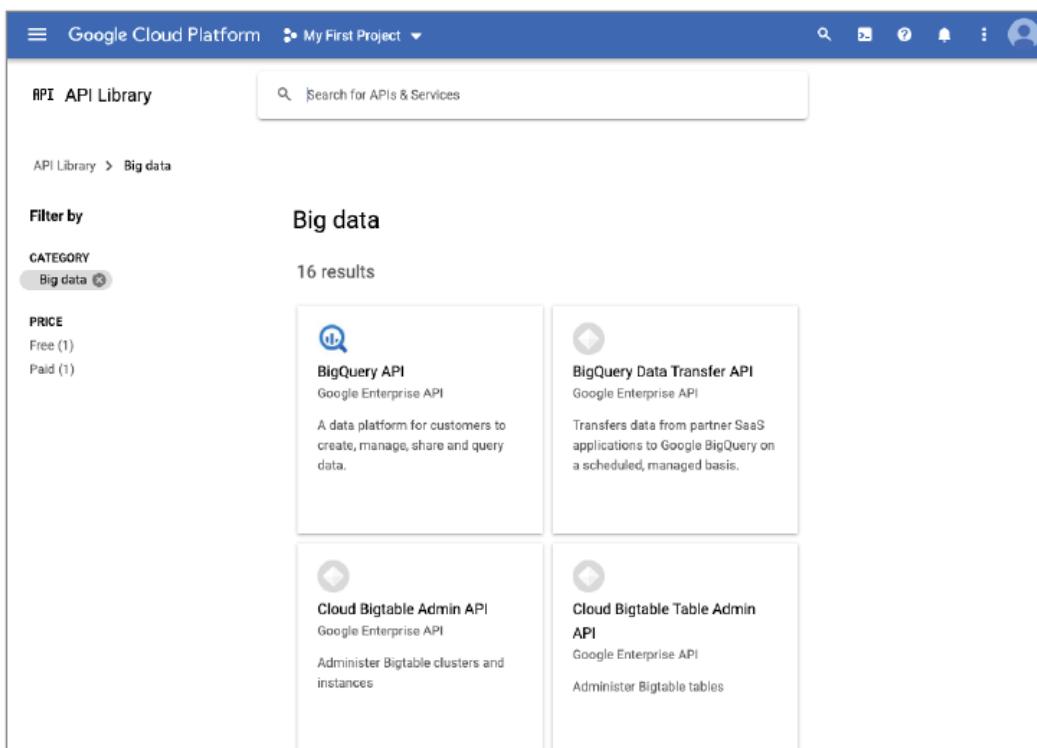
Para habilitar APIs de serviços, você pode selecionar APIs & Serviços no menu principal do console. Isso exibirá um painel, como mostrado na Figura 3.20.

FIGURE 3.20 An example API services dashboard



Se você clicar no link Habilitar APIs e Serviços, verá uma lista de serviços que você pode habilitar, como mostrado na Figura 3.21.

FIGURE 3.21 Example services for Big Data operations



Este formulário é uma maneira conveniente de habilitar APIs de que você sabe que precisará. Se você tentar uma operação que requer uma API que não está habilitada, você pode ser solicitado a decidir se quer habilitar a API.

APIs habilitadas terão uma opção de Desabilitar. Você pode clicar nisso para desativar a API. Você também pode clicar no nome de uma API na lista para obter mais detalhes sobre o uso da API.

Resumo

A abstração central para gerenciar recursos do Google Cloud é a hierarquia de recursos. Ela consiste em três níveis: organização, pasta e projeto. O Serviço de Política de Organização e o IAM juntos controlam o acesso aos recursos de uma organização. Contas de faturamento armazenam informações sobre como pagar por recursos usados. Uma conta de faturamento está associada a um ou mais projetos. O Google Cloud usa APIs para tornar os serviços acessíveis programaticamente e a maioria não está habilitada por padrão.

Essenciais do Exame

Entenda a hierarquia de recursos do Google Cloud. Todos os recursos são organizados dentro de sua hierarquia de recursos. Você pode definir a hierarquia de recursos usando uma organização e várias pastas e projetos. Pastas são úteis para agrupar departamentos e outros grupos gerenciam seus projetos separadamente. Projetos contêm

recursos como VMs e buckets de armazenamento na nuvem. Projetos devem ter contas de faturamento associadas a eles para usar serviços que não são gratuitos.

Entenda as políticas de organização. Políticas de organização restringem recursos na hierarquia de recursos. Políticas incluem restrições, que são regras que definem o que pode ou não pode ser feito com um recurso. Por exemplo, uma restrição pode ser definida para bloquear o acesso à porta serial em todas as VMs em um projeto. Além disso, entenda o processo de avaliação de políticas e como substituir políticas herdadas.

Entenda contas de serviço e como elas são usadas. Contas de serviço são identidades que não estão associadas a um usuário específico, mas podem ser atribuídas a um recurso, como uma VM. Recursos que são atribuídos a uma conta de serviço podem realizar operações que a conta de serviço tem permissão para realizar. Entenda contas de serviço e como criá-las.

Entenda o Faturamento do Google Cloud. O faturamento deve ser habilitado para usar serviços e recursos além dos serviços gratuitos. O faturamento associa um método de faturamento, como um cartão de crédito ou informações de faturamento, a um projeto. Todos os custos associados a recursos em um projeto são cobrados na conta de faturamento do projeto. Uma conta de faturamento pode ser associada a mais de um projeto. Você gerencia seu faturamento por meio da API de Faturamento.

Questões

1. Você está projetando aplicações em nuvem para um provedor de saúde. O aplicativo de gerenciamento de registros gerenciará informações médicas para pacientes. O acesso a esses dados será limitado a um pequeno número de funcionários. O aplicativo do departamento de faturamento terá informações de seguro e pagamento. Outro grupo de funcionários terá acesso às informações de faturamento. Além disso, o sistema de faturamento terá dois componentes: um sistema de faturamento de seguros privados e um sistema de faturamento de pagadores governamentais. Regulamentos governamentais exigem que o software usado para faturar o governo deve ser isolado de outros sistemas de software. Qual das seguintes hierarquias de recursos atenderia a esses requisitos e ofereceria a maior flexibilidade para se adaptar a requisitos em mudança?
 - A. Uma organização, com pastas para gerenciamento de registros e faturamento. A pasta de faturamento teria pastas de seguradora privada e pagador governamental dentro dela. Restrições comuns seriam especificadas em políticas no nível da organização. Outras políticas seriam definidas na pasta apropriada.
 - B. Uma pasta para gerenciamento de registros, uma para faturamento, e nenhuma organização. Políticas definidas no nível da pasta.
 - C. Uma organização, com pastas para gerenciamento de registros, seguradora privada e pagador governamental abaixo da organização. Todas as restrições seriam especificadas em políticas no nível da organização. Todas as pastas teriam as mesmas restrições de política.
 - D. Nenhuma das opções acima.
2. Quando você cria uma hierarquia, você pode ter mais de uma de qual estrutura?
 - A. Apenas organização
 - B. Apenas pasta
 - C. Pasta e projeto
 - D. Apenas projeto
3. Você está projetando um aplicativo que usa uma série de serviços para transformar dados de sua forma original para um formato adequado para uso em um armazém de dados. Seu aplicativo de transformação escreverá na fila de mensagens à medida que processa cada arquivo de entrada. Você não quer dar permissão aos usuários para escrever na fila de mensagens. Você poderia permitir que o aplicativo escrevesse na fila de mensagens usando qual das seguintes opções?
 - A. Conta de faturamento
 - B. Conta de serviço
 - C. Conta de mensagens
 - D. Pasta

4. Sua empresa tem várias políticas que precisam ser aplicadas a todos os projetos. Você decide aplicar políticas à hierarquia de recursos. Pouco depois de aplicar as políticas, um engenheiro descobre que um aplicativo que funcionava antes da implementação das políticas não funciona mais. O engenheiro gostaria que você criasse uma exceção para o aplicativo. Como você pode anular uma política herdada de outra entidade na hierarquia de recursos?
- A. Políticas herdadas podem ser anuladas definindo uma política no nível de pasta ou projeto.
 - B. Políticas herdadas não podem ser anuladas.
 - C. Políticas podem ser anuladas vinculando-as a contas de serviço.
 - D. Políticas podem ser anuladas vinculando-as a contas de faturamento.
5. Restrições são usadas em políticas da hierarquia de recursos. Quais dos seguintes são tipos de restrições permitidas?
- A. Permitir um conjunto específico de valores.
 - B. Negar um conjunto específico de valores.
 - C. Negar um valor e todos os seus valores filhos.
 - D. Permitir todos os valores permitidos.
 - E. Todas as opções acima.
6. Uma equipe com quatro membros quer que você configure um projeto que precisa apenas de permissões gerais para todos os recursos. Você está concedendo a cada pessoa um papel básico para diferentes níveis de acesso, dependendo de suas responsabilidades no projeto. Quais dos seguintes não estão incluídos como papéis básicos no Google Cloud?
- A. Proprietário
 - B. Publicador
 - C. Editor
 - D. Visualizador
7. Você está implantando um novo aplicativo personalizado e deseja delegar algumas tarefas de administração para os engenheiros de DevOps. Eles não precisam de todos os privilégios de um administrador de aplicativo completo, mas precisam de um subconjunto desses privilégios. Que tipo de papel você deve usar para conceder esses privilégios?
- A. Básico
 - B. Predefinido
 - C. Avançado
 - D. Personalizado

8. Quantas organizações você pode criar em uma hierarquia de recursos?
- A. 1
 - B. 2
 - C. 3
 - D. Ilimitado
9. Você foi contatado pelo departamento financeiro da sua empresa para aconselhamento sobre como automatizar pagamentos para os serviços do Google Cloud. Que tipo de conta você recomendaria configurar?
- A. Conta de serviço
 - B. Conta de cobrança
 - C. Conta de recurso
 - D. Conta de crédito
10. Você está experimentando com o Google Cloud para a sua empresa. Você não tem permissão para incorrer em custos. Como você pode experimentar com o Google Cloud sem incorrer em encargos?
- A. Você não pode; todos os serviços incorrem em encargos.
 - B. Você pode usar um cartão de crédito pessoal para pagar os encargos.
 - C. Você pode usar apenas os serviços gratuitos no Google Cloud.
 - D. Você pode usar apenas produtos sem servidor, que são gratuitos para usar.
11. O CFO da sua empresa está preocupado que eles só aprenderão sobre contas de computação em nuvem anormalmente altas depois que as cobranças já tiverem sido incorridas. Que mecanismo no Google Cloud poderia ser usado para abordar a preocupação do CFO?
- A. Monitoramento do Cloud
 - B. Logging do Cloud
 - C. Orçamento e Alertas
 - D. Restrições de Política
12. Uma grande empresa está planejando usar o Google Cloud em várias subdivisões. Cada subdivisão é gerenciada independentemente e tem seu próprio orçamento. A maioria das subdivisões planeja gastar dezenas de milhares de dólares por mês. Como você recomendaria que eles configurarem suas contas de cobrança?
- A. Usar uma única conta de cobrança self-service.
 - B. Usar múltiplas contas de cobrança self-service.
 - C. Usar uma única conta de cobrança faturada.

- D. Usar múltiplas contas de cobrança faturadas.
13. Um aplicativo para uma empresa financeira precisa de acesso a um banco de dados e a um bucket do Cloud Storage. Não existe um papel predefinido que conceda todas as permissões necessárias sem conceder algumas permissões que não são necessárias. Você decide criar um papel personalizado. Ao definir papéis personalizados, você deve seguir qual dos seguintes princípios?
- A. Rotação de deveres
 - B. Princípio da menor quantidade
 - C. Defesa em profundidade
 - D. Menor privilégio
14. Um administrador de aplicativos é responsável por gerenciar todos os recursos em um projeto. Eles querem delegar a responsabilidade por várias contas de serviço para outro administrador. Se contas de serviço adicionais forem criadas, o outro administrador também deverá gerenciá-las. Qual é a melhor maneira de delegar os privilégios necessários para gerenciar as contas de serviço?
- A. Conceder iam.serviceAccountUser ao administrador no nível do projeto.
 - B. Conceder iam.serviceAccountUser ao administrador no nível da conta de serviço.
 - C. Conceder iam.serviceProjectAccountUser ao administrador no nível do projeto.
 - D. Conceder iam.serviceProjectAccountUser ao administrador no nível da conta de serviço.
15. Você trabalha para um varejista com um grande número de lojas. Todas as noites as lojas enviam dados de vendas diárias. Você recebeu a tarefa de criar um serviço que verifica os uploads todas as noites. Você decide usar uma conta de serviço. Seu gerente questiona a segurança da sua solução proposta, particularmente sobre autenticar a conta de serviço. Você explica o mecanismo de autenticação usado pelas contas de serviço. Qual mecanismo de autenticação é usado?
- A. Nome de usuário e senha
 - B. Autenticação de dois fatores
 - C. Chaves de criptografia
 - D. Biometria
16. Quais objetos no Google Cloud são às vezes tratados como recursos e às vezes como identidades?
- A. Contas de faturamento
 - B. Contas de serviço
 - C. Projetos
 - D. Papéis

17. Você planeja desenvolver uma aplicação web usando produtos do Google Cloud que já incluem papéis estabelecidos para gerenciar permissões, como acesso somente leitura ou a habilidade de deletar versões antigas. Qual dos seguintes papéis oferece essas capacidades?
- A. Papéis Básicos
 - B. Papéis Predefinidos
 - C. Papéis Personalizados
 - D. Papéis de Aplicação
18. Você está revisando uma nova conta do Google Cloud criada para uso pelo departamento financeiro. Um auditor tem perguntas sobre quem pode criar projetos por padrão. Você explica quem tem privilégios para criar projetos por padrão. Quem está incluído?
- A. Apenas administradores de projeto
 - B. Todos os usuários
 - C. Apenas usuários sem o papel resourcemanager.projects.create
 - D. Apenas usuários de conta de faturamento
19. Quantos projetos podem ser criados em uma conta?
- A. 10.
 - B. 25.
 - C. Não há limite.
 - D. Cada conta tem um limite determinado pelo Google.
20. Você está planejando como conceder privilégios aos usuários da conta do Google Cloud da sua empresa. Você precisa documentar o que cada usuário será capaz de fazer. Auditores estão mais preocupados com um papel chamado Administrador da Organização. Você explica que usuários com esse papel podem realizar várias tarefas, que incluem todas as seguintes, exceto qual?
- A. Definir a estrutura da hierarquia de recursos
 - B. Determinar quais permissões um usuário deve ser atribuído
 - C. Definir políticas IAM sobre a hierarquia de recursos
 - D. Delegar outros papéis de gerenciamento para outros usuários

Capítulo 4

Introdução à Computação no Google Cloud

ESTE CAPÍTULO COBRE O SEGUINTE OBJETIVO DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 2.2 Planejamento e configuração de recursos de computação

Neste capítulo, você aprenderá sobre cada uma das opções de computação disponíveis no Google Cloud e quando usá-las. Também discutiremos máquinas virtuais preemptivas e quando elas podem ajudar a reduzir seus custos gerais de computação.

Compute Engine

Compute Engine é um serviço que fornece máquinas virtuais (VMs) que rodam no Google Cloud. Geralmente nos referimos a uma VM em execução como uma instância. Ao usar o Compute Engine, você cria e gerencia uma ou mais instâncias.

Imagens de Máquina Virtual

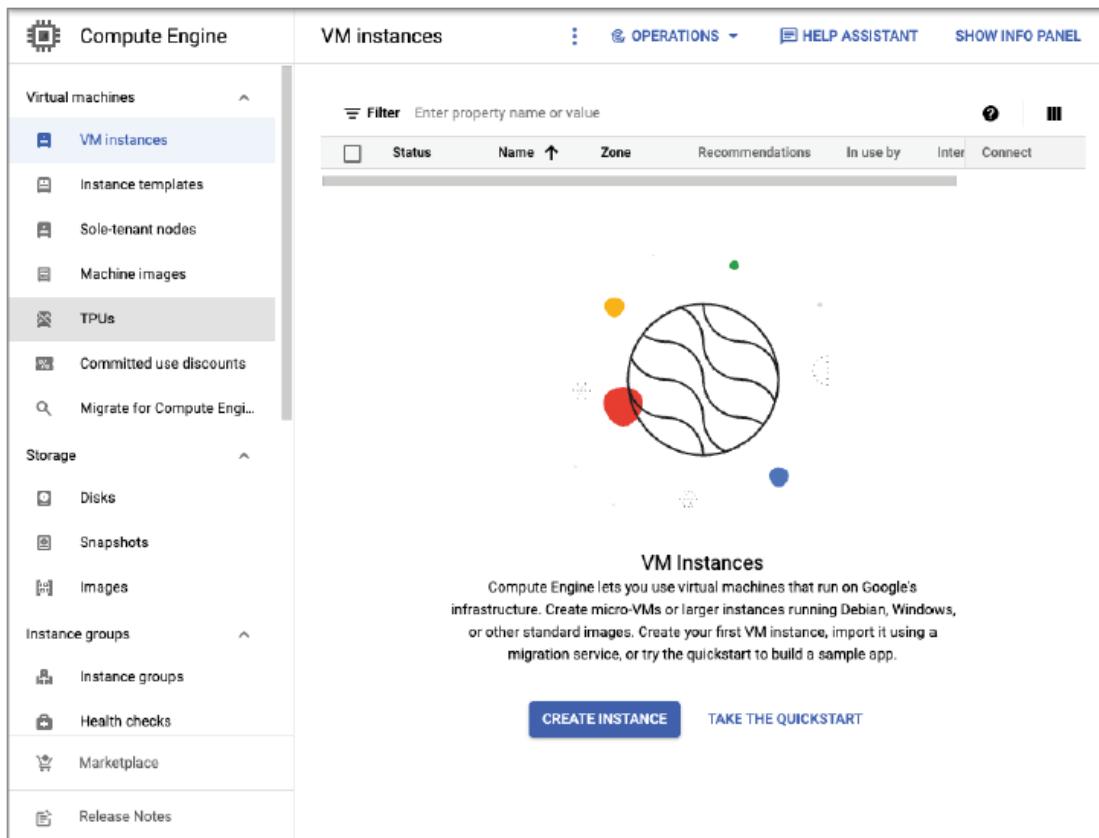
As instâncias executam imagens, que contêm sistemas operacionais, bibliotecas e outros códigos. Você pode escolher executar uma imagem pública fornecida pelo Google (Figura 4.1). Imagens do Linux e do Windows estão disponíveis. Além das imagens mantidas pelo Google, há outras imagens públicas fornecidas por projetos de código aberto ou fornecedores terceirizados.

FIGURA 4.1 Um subconjunto de imagens de sistema operacional disponíveis no Compute Engine

Compute Engine		Images	[+] CREATE IMAGE	REFRESH	DELETE	HELP ASSISTANT	SHOW INFO PANEL		
Virtual machines	^	v20220202-debian-10							
VM instances		<input type="checkbox"/> <input checked="" type="checkbox"/>	c2-deeplearning-pytorch-1-10-cu110-v20220202-debian-10	asia, eu, us	—	50 GB	Debian	pytorch-1-10-gpu-debian-10	Feb 2, 2022, 1:34:16 PM UTC-08:00
Instance templates		<input type="checkbox"/> <input checked="" type="checkbox"/>	c2-deeplearning-pytorch-1-10-xla-v20220202-debian-10	asia, eu, us	—	50 GB	Debian	pytorch-1-10-xla-debian-10	Feb 2, 2022, 12:27:53 PM UTC-08:00
Sole-tenant nodes		<input type="checkbox"/> <input checked="" type="checkbox"/>	centos-7-v20200403	asia, eu, us	—	20 GB	CentOS	centos-7	Apr 6, 2020, 2:51:35 PM UTC-07:00
Machine images		<input type="checkbox"/> <input checked="" type="checkbox"/>	centos-7-v20220126	asia, eu, us	—	20 GB	CentOS	centos-7	Jan 26, 2022, 2:27:27 PM UTC-08:00
TPUs		<input type="checkbox"/> <input checked="" type="checkbox"/>	centos-stream-8-v20220128	asia, eu, us	—	20 GB	CentOS	centos-stream-8	Jan 28, 2022, 11:16:25 AM UTC-08:00
Committed use discounts		<input type="checkbox"/> <input checked="" type="checkbox"/>	cos-69-10895-385-0	asia, eu, us	—	10 GB	Google	cos-69-lts	Oct 8, 2019, 11:25:22 PM UTC-07:00
Migrate for Compute Eng...		<input type="checkbox"/> <input checked="" type="checkbox"/>	cos-73-11647-656-0	asia, eu, us	—	10 GB	Google	cos-73-lts	Sep 5, 2020, 2:25:50 PM UTC-07:00
Storage	^	<input type="checkbox"/> <input checked="" type="checkbox"/>	cos-77-12371-1109-0	asia, eu, us	—	10 GB	Google	cos-77-lts	Jan 11, 2021, 11:36:57 AM UTC-08:00
Disks									
Snapshots									
Images									
Instance groups	^								
Instance groups									
Health checks									
Marketplace									
Release Notes									

As imagens públicas incluem uma variedade de sistemas operacionais, como CentOS, Container-Optimized OS da Google, Debian, Red Hat Enterprise Linux, SUSE Enterprise Linux Server, Ubuntu e Windows Server. Caso não haja uma imagem pública que atenda às suas necessidades, você pode criar uma imagem personalizada a partir de um disco de inicialização ou começando com outra imagem. Para criar uma VM a partir do console, navegue até Compute Engine e depois até VM Instances. Você verá uma tela similar à Figura 4.2.

FIGURE 4.2 Creating a VM in Compute Engine



Clique em Criar Instância para abrir a página de criação de uma instância. Aqui, como mostrado na Figura 4.3, você pode definir o nome da instância, escolher a configuração da máquina, adicionar unidades de processamento gráfico (GPUs) e definir outras características da instância.

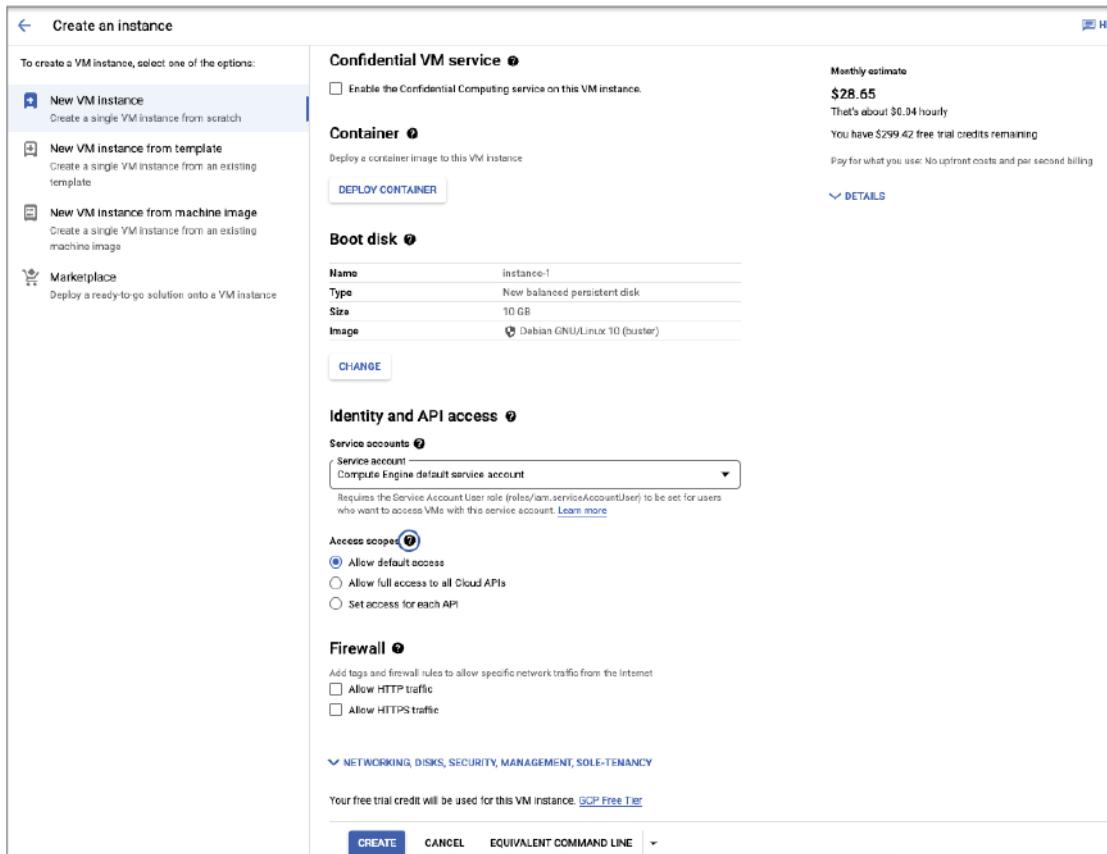
Outras características configuráveis de uma instância são mostradas na Figura 4.4. Por exemplo, para aplicações de alta segurança, você pode usar o serviço Confidential VM para criptografar dados na memória. Você também pode especificar um nome, tamanho, imagem e tipo do disco de inicialização. VMs têm uma identidade associada chamada de conta de serviço associada a elas. Contas de serviço são identidades, como usuários e grupos, mas não estão associadas a usuários humanos. Contas de serviço podem ser atribuídas a papéis para que possam ter permissões para realizar ações no Google Cloud. (Para mais sobre contas de serviço, veja o Capítulo 3, "Projetos, Contas de Serviço e Cobrança").

FIGURE 4.3 Part 1 of creating an instance in Compute Engine

Você também pode controlar quais ações uma instância pode executar configurando escopos de acesso. Escopos de acesso são um mecanismo de controle de acesso legado que existia antes do serviço de Gerenciamento de Identidade e Acesso (IAM). Por padrão, os escopos de acesso permitem acesso mínimo, incluindo a capacidade de ler do armazenamento e escrever para serviços de monitoramento e registro. IAM é o método preferido para controlar o acesso concedido a uma instância do Compute Engine.

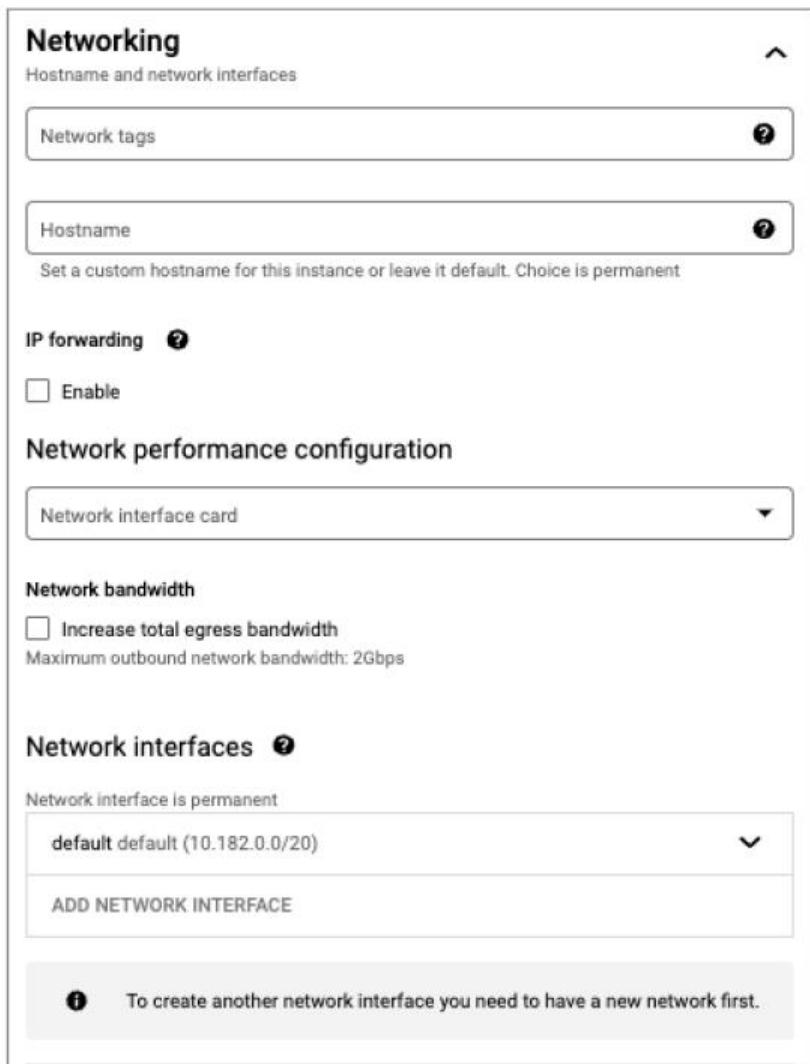
Você também pode especificar se o tráfego HTTP ou HTTPS é permitido para a instância. A Figura 4.5 mostra as configurações de rede para uma instância. Na seção de Rede da página Criar Instância, você pode especificar tags de rede, um nome de host e configurações de desempenho de rede, bem como adicionar interfaces de rede adicionais. Uma interface de rede é criada por padrão.

FIGURE 4.4 Part 2 of creating an instance in Compute Engine



Se você deseja discos adicionais, juntamente com o disco de inicialização, você também pode adicionar e configurar discos nesta página. A Figura 4.6 mostra as opções para configurar discos. Você pode fornecer um nome, descrição, tipo de disco, tamanho, um cronograma de backup para o disco e configurações de criptografia. Note que todos os dados no Google Cloud são criptografados quando armazenados (conhecido como criptografia em repouso) em armazenamento persistente. Não temos a opção de armazenar dados de forma persistente sem criptografia, mas podemos escolher como as chaves de criptografia são gerenciadas. Atualmente, as escolhas são chaves de criptografia gerenciadas pelo Google, chaves de criptografia gerenciadas pelo cliente e chaves de criptografia fornecidas pelo cliente. Com chaves de criptografia gerenciadas pelo Google, o Google cria e gerencia as chaves. Com chaves de criptografia fornecidas pelo cliente, os clientes criam suas próprias chaves, mas o Google as gerencia. Quando usamos chaves de criptografia fornecidas pelo cliente, os clientes criam e gerenciam as chaves fora do Google Cloud.

FIGURE 4.5 Configuring network properties in a Compute Engine instance



Os discos podem ser anexados como discos de leitura/escrita ou como disco de somente leitura. Por padrão, um disco é mantido quando uma instância é deletada, mas você pode escolher ter o disco deletado quando a instância for deletada. Na seção de Segurança da página Criar Instância, você pode especificar algumas características de segurança avançadas. (Veja a Figura 4.7.) O Secure Boot protege contra códigos maliciosos no nível de inicialização e no nível do kernel, como rootkits. O Módulo de Plataforma Confiável Virtual (vTPM) valida a integridade da inicialização e fornece proteções adicionais para a geração e proteção de chaves. Quando o vTPM está habilitado, você tem a opção de habilitar o Monitoramento de Integridade, que verifica a integridade de execução da máquina virtual.

FIGURE 4.6 Configuring disks in a Compute Engine instance

Add new disk X

Name * ?
Name is permanent

Description

Source
Create a blank disk, apply a bootable disk image, or restore a snapshot of another disk in this project.

Disk source type * ?

Disk settings

Disk type * ?

[COMPARE DISK TYPES](#)

Size * GB ?
Provision between 10 and 65,536 GB

Snapshot schedule (Recommended)
Use snapshot schedules to automate disk backups. [Learn more](#)

Select a snapshot schedule ▼

Encryption
Data is encrypted automatically. Select an encryption key management solution.

Google-managed encryption key
No configuration required

Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

Customer-supplied encryption key (CSEK)
Manage outside of Google Cloud

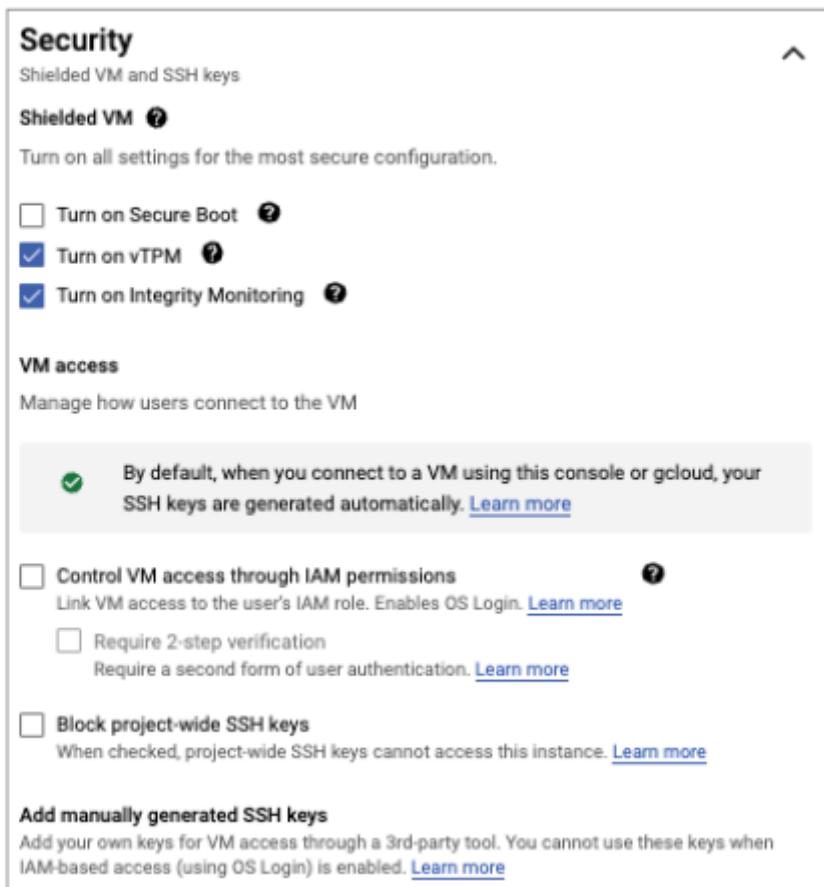
Labels ?
[+ ADD LABEL](#)

Attachment settings

Mode

SAVE CANCEL

FIGURE 4.7 Configuring security in a Compute Engine instance



Você pode restringir ainda mais o acesso a uma instância por meio de papéis do IAM. Quando esse recurso está habilitado, apenas usuários com o papel de Login de SO do Compute, Login de Admin do SO do Compute ou outros papéis que tenham permissões para habilitar acesso baseado em IAM podem fazer login. Outra maneira de bloquear o acesso é proibir o uso de chaves SSH baseadas em projetos, que por padrão permitiriam acesso a qualquer instância de VM em um projeto. A Figura 4.8 mostra as opções para especificar recursos de gerenciamento. Estes incluem uma descrição, a capacidade de bloquear a exclusão da instância, reservas de instância (uma maneira de comprar blocos de tempo de instância com desconto) e se você deseja que um script de automação seja executado na inicialização. Você também pode configurar parâmetros de disponibilidade, incluindo optar por tornar esta uma VM preemptiva. VMs preemptivas custam menos, mas podem ser desligadas a qualquer momento pelo Google Cloud. Originalmente, as VMs preemptivas rodariam por no máximo 24 horas antes de serem desligadas. Agora, o Google Cloud oferece VMs spot, que são cobradas como VMs preemptivas, mas não são necessariamente desligadas após 24 horas. Você também pode especificar se a instância deve ser migrada para outro servidor durante a manutenção do servidor e iniciada automaticamente se houver uma falha de hardware ou outro desligamento não iniciado pelo usuário.

FIGURE 4.8 Configuring management features in a Compute Engine instance

Management

Description, deletion protection, reservations, automation, and availability policies

Description

Deletion protection ?

Enable deletion protection

Reservations

Automatically use created reservation

Use an existing reservation when creating this VM instance

Automation

Startup script

You can choose to specify a startup script that will run when your instance boots up or restarts. Startup scripts can be used to install software and updates, and to ensure that services are running within the virtual machine. [Learn more](#)

Metadata

You can set custom metadata for an instance or project outside of the server-defined metadata. This is useful for passing in arbitrary values to your project or instance that can be queried by your code on the instance. [Learn more](#)

+ ADD ITEM

Availability policy

Preemptibility

Off (Recommended)

A preemptible VM costs much less, but lasts only 24 hours. It can be terminated sooner due to system demands. [Learn more](#)

On host maintenance

Migrate VM instance (Recommended)

When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart

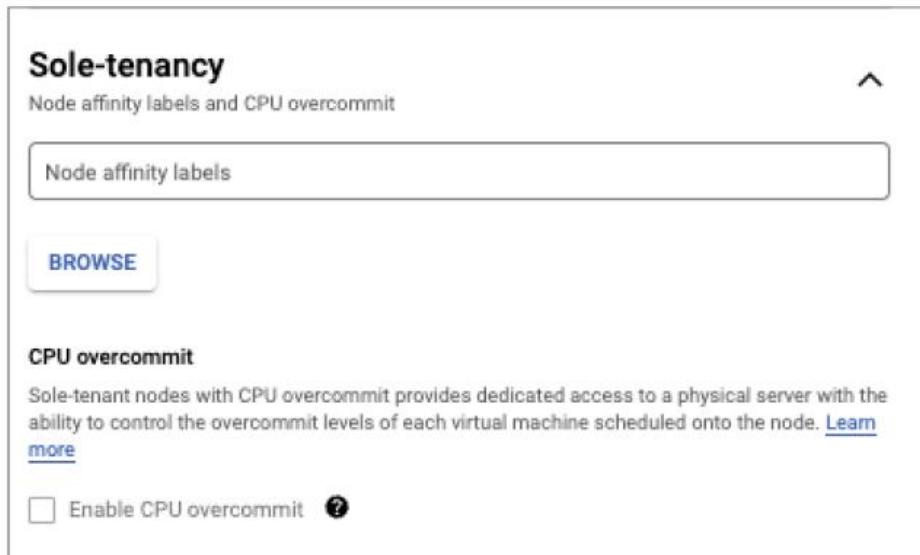
On (recommended)

Compute Engine can automatically restart VM instances if they are terminated for non-user-initiated reasons (maintenance event, hardware failure, software failure and so on)

Pode haver momentos em que você não deseja que máquinas virtuais de outros projetos sejam executadas no mesmo servidor que as máquinas virtuais do seu projeto. Nesses casos, você pode escolher a Locação Exclusiva para sua instância (veja a Figura 4.9). Apenas VMs do seu projeto com etiquetas de afinidade de nó correspondendo às etiquetas que você especificar aqui serão executadas no mesmo servidor juntas. Você também tem a opção de sobrecarregar as CPUs em um servidor configurado como inquilino único. Isso pode aumentar o desempenho ao agendar VMs com mais requisitos

de CPU do que realmente disponíveis se as VMs não precisarem de todos os recursos comprometidos ao mesmo tempo. Por exemplo, se duas instâncias estão rodando em um servidor e uma instância tem um pico de carga pela manhã e a outra tem um pico de carga à noite, você pode ser capaz de sobrecarregar sem impactar negativamente o desempenho de qualquer instância.

FIGURE 4.9 Configuring Sole Tenancy features in a Compute Engine instance



Se você vai criar instâncias adicionais com a mesma configuração, você pode criar um modelo de instância. Um modelo é uma descrição de uma configuração de VM. O processo de criar um modelo de instância é similar a criar uma VM como acabou de ser descrito, mas em vez de criar uma VM ao completar, você terá criado um modelo. Você pode então usar esse modelo para criar uma nova instância sem ter que especificar todos os parâmetros de configuração manualmente. Outra maneira de criar uma instância é a partir de uma imagem de máquina que você cria. A Figura 4.10 mostra a caixa de diálogo para criar uma imagem de máquina a partir de uma VM existente. Você especifica um nome, descrição, VM de origem e local para armazenar a imagem. Você também pode especificar como as chaves de criptografia são gerenciadas.

FIGURE 4.10 Creating a machine image

Create a machine image

A machine image contains a VM's properties, metadata, permissions, and data from all its attached disks. You can use a machine image to create, backup, or restore a VM.
[Learn more](#)

Name *

Name is permanent

Description

Source VM instance *

Location

Multi-regional
 Regional

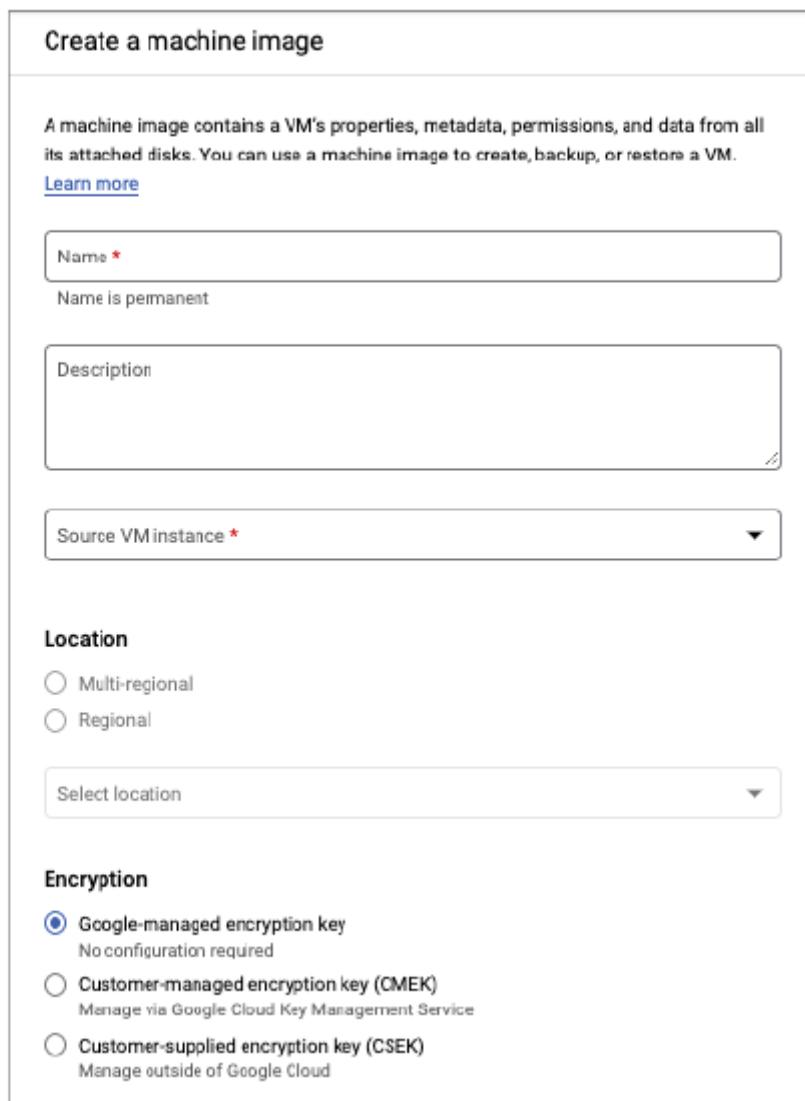
Select location ▾

Encryption

Google-managed encryption key
No configuration required

Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

Customer-supplied encryption key (CSEK)
Manage outside of Google Cloud

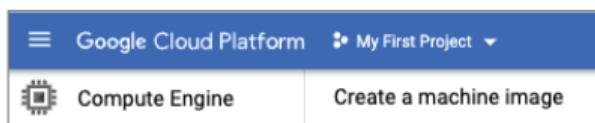


As Máquinas Virtuais Estão Contidas em Projetos

Quando você cria uma instância, você especifica um projeto para conter a instância. Como você pode se lembrar, projetos fazem parte da hierarquia de recursos do Google Cloud. Projetos são a estrutura de nível mais baixo na hierarquia. Projetos permitem que você gerencie recursos relacionados com políticas comuns.

Quando você abre o Google Cloud Console, você notará no topo do formulário o nome de um projeto ou a frase Seleciona Um Projeto, como mostrado na Figura 4.11.

FIGURE 4.11 The current project name or the option to select one is displayed in Google Cloud Console.



Quando você escolhe Selecionar Um Projeto, um formulário como o da Figura 4.12 aparece. A partir daí, você pode selecionar o projeto onde quer armazenar seus recursos, incluindo VMs.

FIGURE 4.12 Choosing a project from existing projects in an account



Máquinas Virtuais Rodam em uma Zona e Região

Além de ter um projeto, instâncias de VM têm uma zona atribuída. Zonas são recursos semelhantes a centros de dados, mas podem consistir de um ou mais centros de dados estreitamente acoplados. Eles estão localizados dentro de regiões. Uma região é uma localização geográfica, como asia-east1, europe-west2 e us-east4.

As zonas dentro de uma região são ligadas por conexões de rede de alta largura de banda e baixa latência.

Você especifica uma região e uma zona quando cria uma VM. Como você pode ver na Figura 4.13 e Figura 4.14, o formulário Criar VM inclui listas suspensas das quais você pode selecionar a região e a zona.

Você pode querer considerar vários fatores ao escolher onde executar sua VM:

- Custo, que pode variar entre regiões.
- Regulações de localidade de dados, como manter dados sobre cidadãos da União Europeia na União Europeia.
- Alta disponibilidade. Se você estiver executando múltiplas instâncias, você pode querê-las em diferentes zonas e possivelmente diferentes regiões. Se uma das zonas ou regiões se tornar inacessível, as instâncias em outras zonas e regiões ainda podem fornecer serviços.

FIGURE 4.13 Selecting a region in the Create VM form

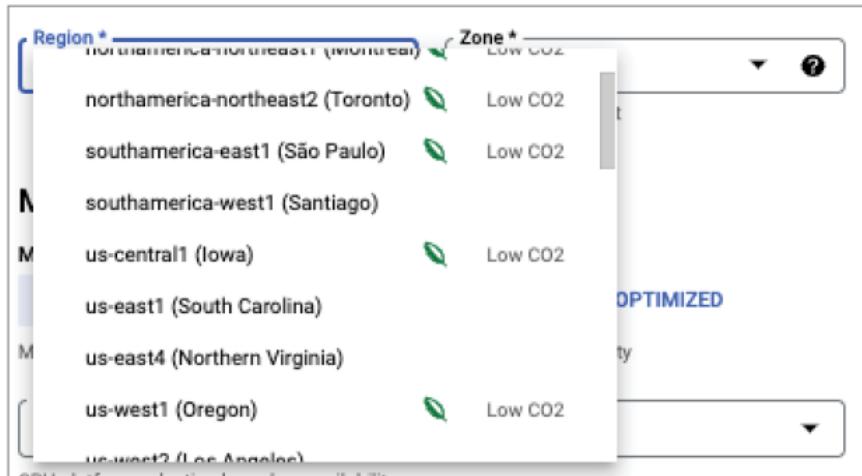
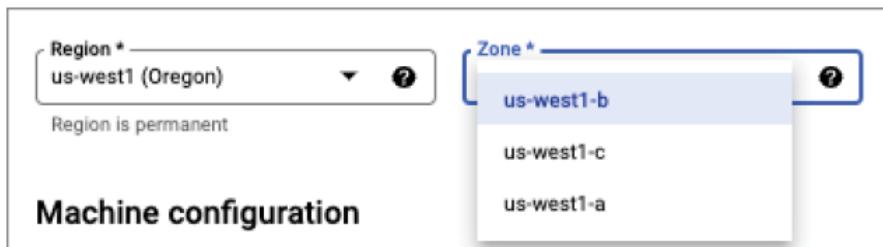


FIGURE 4.14 Once a region is selected, you can choose a zone within that region.



- Latência, que é importante se você tem usuários em diferentes partes do mundo. Manter instâncias e dados geograficamente próximos aos usuários da aplicação pode ajudar a reduzir a latência.

- Necessidade de plataformas de hardware específicas, que podem variar por região. Por exemplo, europe-west1 pode ter um processador disponível que não está disponível em europe-west2.

- A intensidade de carbono da geração de energia na região.

Usuários Precisam de Privilégios para Criar Máquinas Virtuais

Para criar recursos do Compute Engine em um projeto, usuários devem ser membros do projeto ou de um recurso específico e ter permissões apropriadas para realizar tarefas específicas. Usuários podem ser associados a projetos da seguinte forma:

- Usuários individuais
- Um grupo Google
- Um domínio do Google Workspace
- Uma conta de serviço

Uma vez que um usuário, ou um conjunto de usuários, é adicionado a um projeto, você pode atribuir permissões concedendo papéis ao usuário ou conjunto de usuários. Esse processo é explicado em detalhe no Capítulo 17, "Configurando Acesso e

Segurança". Os papéis predefinidos são especialmente úteis porque eles agrupam permissões que são frequentemente necessárias para um usuário realizar um conjunto de tarefas. Aqui estão alguns exemplos de papéis predefinidos:

Usuários Administradores de Computação com este papel têm controle total sobre as instâncias do Compute Engine.

Usuários Administradores de Rede de Computação com este papel podem criar, modificar e excluir a maioria dos recursos de rede, e proporcionam acesso somente leitura a regras de firewall e certificados SSL. Esse papel não dá ao usuário permissão para criar ou alterar instâncias.

Usuários Administradores de Segurança de Computação com este papel podem criar, modificar e excluir certificados SSL e regras de firewall.

Usuários Visualizadores de Computação com este papel podem obter e listar recursos do Compute Engine, mas não podem ler dados desses recursos.

Quando privilégios são concedidos a usuários no nível do projeto, então essas permissões aplicam-se a todos os recursos dentro de um projeto. Por exemplo, se um usuário recebe o papel de Administrador de Computação no nível do projeto, então essa pessoa pode administrar todas as instâncias do Compute Engine no projeto.

Uma maneira alternativa de controlar o acesso aos recursos é anexar políticas de IAM diretamente aos recursos. Dessa forma, privilégios podem ser personalizados para recursos específicos em vez de para todos os recursos em um projeto. Por exemplo, você poderia especificar que a usuária Alice tem o papel de Administrador de Computação em uma instância e Bob tem o mesmo papel em outra instância. Alice e Bob seriam capazes de administrar suas próprias instâncias de VM, mas não poderiam administrar outras instâncias.

Máquinas Virtuais Preemptíveis

Considere se você tem uma carga de trabalho que é o oposto de precisar de alta disponibilidade. VMs preemptíveis são instâncias de computação de curta duração adequadas para executar certos tipos de cargas de trabalho—particularmente para aplicações que realizam modelagem financeira, renderização, big data, integração contínua e operações de web crawling. Estas VMs oferecem as mesmas opções de configuração que instâncias regulares de computação e persistem por até 24 horas; VMs spot não têm esta limitação de tempo. Se uma aplicação é tolerante a falhas e pode suportar possíveis interrupções de instâncias (com um aviso de 30 segundos), então o uso de instâncias de VM preemptíveis e VMs spot pode reduzir significativamente os custos do Google Compute Engine.

Alguns trabalhos de análise de big data são executados em clusters de servidores rodando softwares como Hadoop e Spark. As plataformas são projetadas para serem resilientes a falhas. Se um nó cai no meio de um trabalho, a plataforma detecta a falha e move a carga de trabalho para outros nós no servidor. Você pode ter trabalhos analíticos que são bem atendidos por uma combinação de VMs confiáveis e VMs preemptíveis.

Com alguma porcentagem de VMs confiáveis, você sabe que pode ter seus trabalhos processados dentro das suas restrições de tempo, mas se você adicionar VMs preemptíveis de baixo custo, você pode muitas vezes terminar seus trabalhos mais rápido e com menor custo total.

Limitações das Máquinas Virtuais Preemptíveis

Ao decidir onde usar VMs preemptíveis, tenha em mente suas limitações e diferenças comparadas às instâncias convencionais de VM no Google Cloud. VMs preemptíveis têm as seguintes características:

- Podem ser terminadas a qualquer momento. Se forem terminadas dentro de 1 minuto após o início, você não será cobrado por esse tempo.
- Serão terminadas dentro de 24 horas, exceto para VMs spot.
- Podem não estar sempre disponíveis. A disponibilidade pode variar entre zonas e regiões.
- Não podem migrar para uma VM regular.
- Não podem ser configuradas para reiniciar automaticamente.
- Não estão cobertas por nenhum acordo de nível de serviço (SLA).

Tipos de Máquinas Personalizadas

O Compute Engine possui dezenas de tipos de máquinas predefinidos agrupados em tipos padrão, máquinas de alta memória, máquinas de alto CPU, tipo de núcleo compartilhado e máquinas otimizadas para memória. Esses tipos de máquinas predefinidos variam no número de CPUs virtuais (vCPUs) e quantidade de memória. Aqui estão alguns exemplos:

- n2-standard-2 tem 2 vCPU e 8 GB de memória.
- n2-standard-32 tem 32 vCPUs e 128 GB de memória.
- m2-megamem-416 tem 416 vCPUs e 5,75 TB de memória.
- m2-ultramem-208 tem 208 vCPUs e 5,75 TB de memória.

As opções predefinidas para VMs atenderão às necessidades de muitos casos de uso, mas pode haver momentos em que sua carga de trabalho poderia ser executada de maneira mais custo-efetiva e mais rápida em uma configuração que não está já definida. Nesse caso, você pode querer usar um tipo de máquina personalizado.

Para criar uma imagem personalizada, selecione a opção Criar VM no console. Clique no link Personalizar na seção Tipo de Máquina (veja a Figura 4.15).

Isso expande a seção Tipo de Máquina, como mostrado na Figura 4.16. A partir daí, você pode ajustar os controles deslizantes para aumentar ou diminuir o número de CPUs e a quantidade de memória que você precisa.

As opções disponíveis para criar uma configuração de máquina personalizada variarão por série. Por exemplo, tipos de máquinas personalizadas baseadas na série N2 podem ter entre 2 e 80 vCPUs e até 640 GB de memória. O preço de uma configuração personalizada é baseado no número de vCPUs e na memória alocada. Tipos de máquinas personalizadas baseadas na série N2D podem ter até 96 núcleos e até 768 GB de memória. Você pode selecionar Estender Memória para aumentar a quantidade de memória em relação às CPUs.

FIGURE 4.15 Choosing a custom machine type from the MachineType drop-down menu

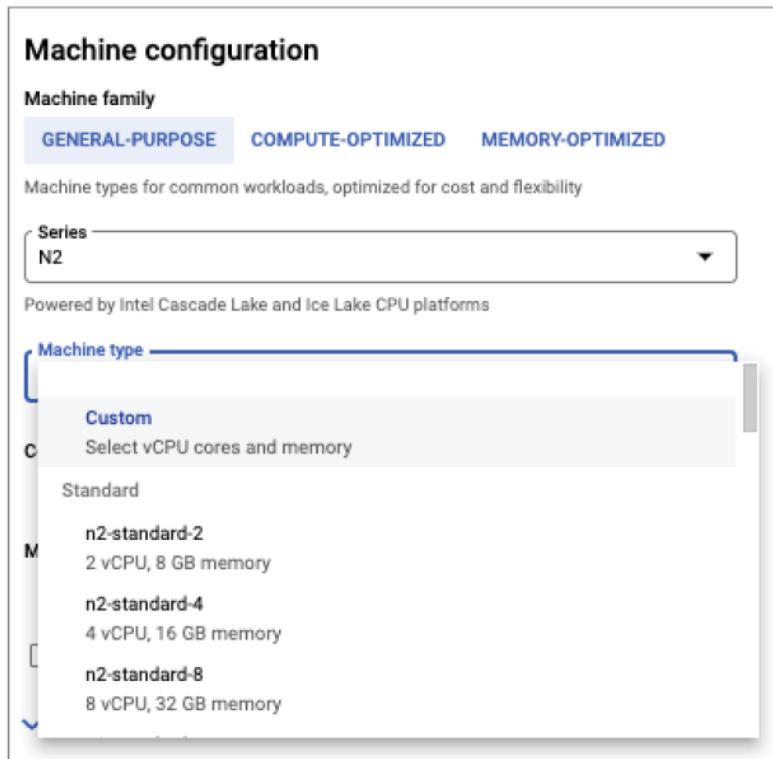
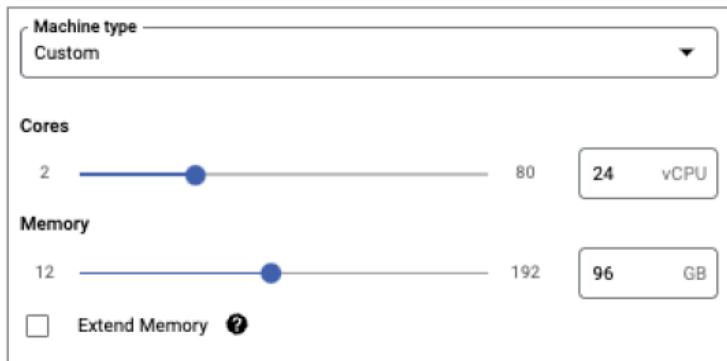


FIGURE 4.16 Customizing a VM by adjusting the number of CPUs and the amount of memory



Casos de Uso para Máquinas Virtuais do Compute Engine

O Compute Engine é uma boa opção quando você precisa de controle máximo sobre instâncias de VM. Com o Compute Engine, você pode fazer o seguinte:

- Escolher a imagem específica para executar na instância.
- Instalar pacotes de software ou bibliotecas personalizadas.
- Ter controle detalhado sobre quais usuários têm permissões na instância.
- Ter controle sobre certificados SSL e regras de firewall para a instância.

Em relação a outros serviços de computação no Google Cloud, o Google Compute Engine oferece o menor nível de gerenciamento. O Google fornece imagens públicas e

um conjunto de configurações de VM, mas você, como administrador, deve fazer escolhas sobre qual imagem usar, o número de CPUs, a quantidade de memória para alocar, como configurar o armazenamento persistente e como configurar configurações de rede.

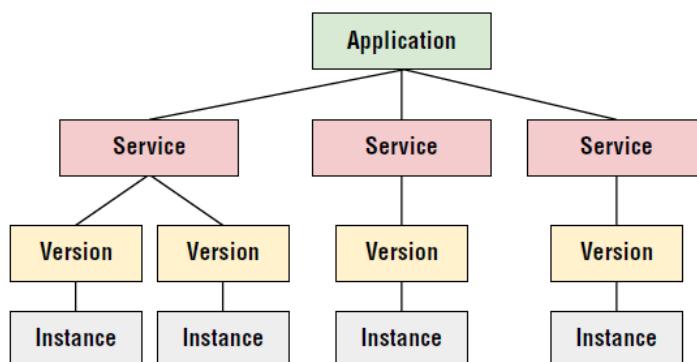
Em geral, quanto mais controle sobre um recurso você tem no Google Cloud, mais responsabilidade você tem pela configuração e gerenciamento do recurso.

App Engine

O App Engine é um serviço de computação PaaS que fornece uma plataforma gerenciada para executar aplicações. Quando você usa o App Engine, seu foco está na sua aplicação e não nas VMs que executam a aplicação. Em vez de configurar VMs, você especifica alguns requisitos básicos de recursos junto com seu código de aplicação, e o Google gerenciará os recursos necessários para executar o código. Isso significa que os usuários do App Engine têm menos a gerenciar, mas também têm menos controle sobre os recursos de computação que são usados para executar a aplicação.

Como as instâncias de VM, as aplicações no App Engine são criadas dentro de um projeto. Diferente do Compute Engine, ao criar um serviço do App Engine, você não está fornecendo muitos detalhes para configurar máquinas virtuais. Em vez disso, você está configurando sua aplicação para rodar como um serviço no App Engine (veja a Figura 4.17).

FIGURE 4.17 When using App Engine, the focus is on applications, not infrastructure.

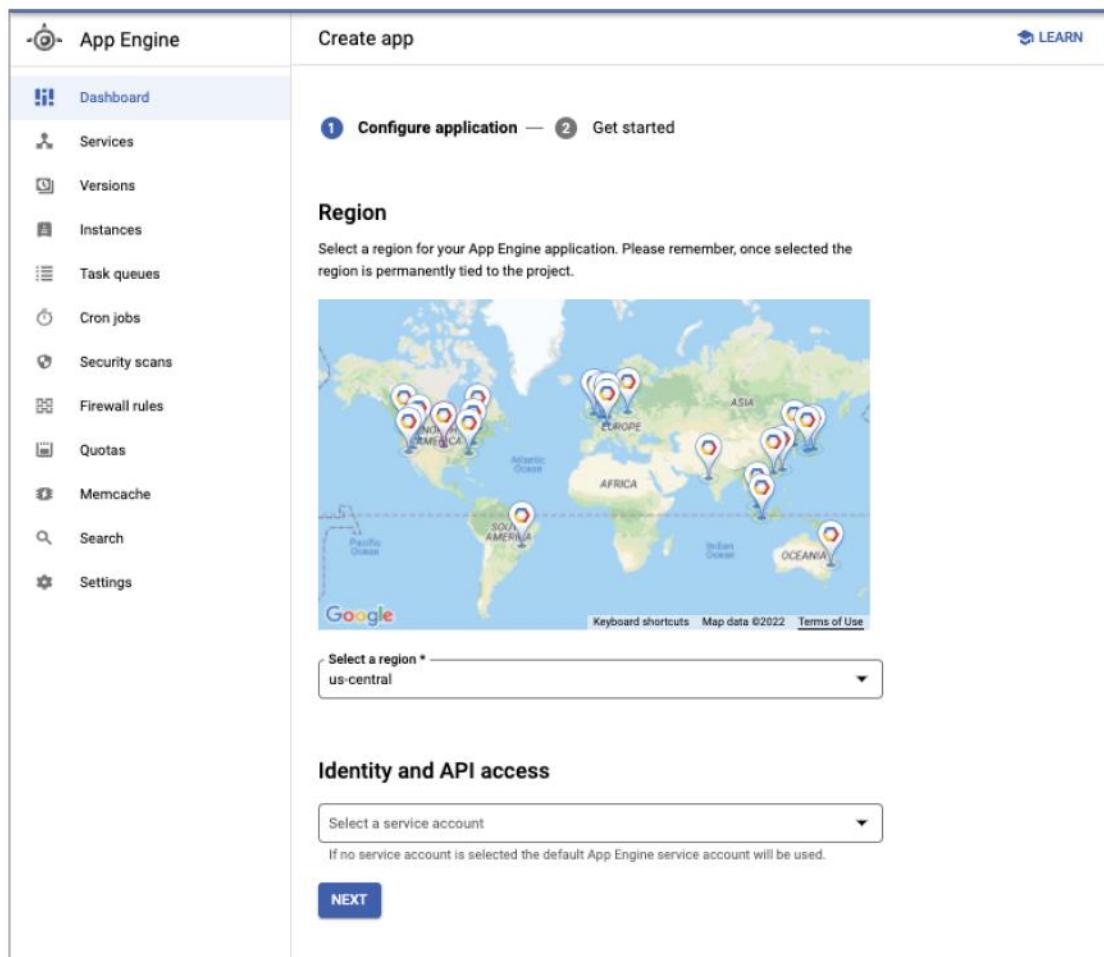


O App Engine não está incluído como tópico no guia do exame Google Cloud Associate Cloud Engineer, mas é incluído aqui porque o serviço ainda está disponível e continua sendo usado.

Estrutura de uma Aplicação do App Engine

Aplicações do App Engine têm uma estrutura comum e consistem em serviços. Serviços fornecem uma função específica, como calcular o imposto sobre vendas em uma aplicação web de varejo ou atualizar o inventário conforme produtos são vendidos em um site. Serviços têm versões, e isso permite que múltiplas versões rodam ao mesmo tempo. Cada versão de um serviço roda em uma instância que é gerenciada pelo App Engine (veja a Figura 4.18).

FIGURE 4.18 The structure of an App Engine application



O número de instâncias usadas para fornecer uma aplicação depende da sua configuração para a aplicação e da carga atual sobre a aplicação. Conforme a carga aumenta, o Google pode adicionar mais instâncias para atender à necessidade. Da mesma forma, se a carga diminui, instâncias podem ser desligadas para economizar no custo de instâncias não utilizadas. Esse tipo de autoescalonamento está disponível com instâncias dinâmicas.

Além de instâncias dinâmicas, o App Engine também fornece instâncias residentes. Você pode adicionar ou remover instâncias residentes manualmente.

Quando o número de instâncias implantadas muda frequentemente, pode ser difícil estimar os custos de execução das instâncias. Felizmente, o Google Cloud permite aos usuários configurar limites diários de gastos, bem como criar orçamentos e configurar alarmes.

Ambientes Padrão e Flexível do App Engine

O App Engine fornece dois tipos de ambientes de execução: padrão e flexível. O ambiente padrão fornece ambientes de execução de linguagem, enquanto o ambiente flexível é uma plataforma de execução de contêiner mais generalizada. Em ambos os ambientes, seu código roda em instâncias de contêiner executadas na infraestrutura gerenciada pelo Google Cloud.

Ambiente Padrão do App Engine

O ambiente padrão é o ambiente original do App Engine. Consiste em um ambiente de execução específico para cada linguagem e pré-configurado. Atualmente, existem duas gerações do ambiente padrão. A segunda geração melhora o desempenho da primeira geração e tem menos limitações.

Atualmente, os usuários do ambiente padrão do App Engine podem escolher entre as seguintes linguagens suportadas:

Primeira Geração

- Python 2.7
- Java 8
- PHP 5.5
- Go 1.11

Segunda Geração

- Python 3
- Java 11, 17
- Node.js
- PHP 7/8
- Ruby
- Go 1.12+

Com o ambiente padrão de segunda geração, os desenvolvedores podem usar qualquer extensão de linguagem, mas na primeira geração apenas um conjunto selecionado de extensões e bibliotecas é permitido. O acesso à rede é restrito na primeira geração, mas os usuários têm acesso total à rede na segunda geração.

Os serviços do App Engine são escalados usando escalonamento automático, manual ou básico. Com o escalonamento básico, o App Engine tenta manter os custos baixos, então ele não inicia outra instância até que haja uma solicitação que não possa ser atendida por uma instância existente. Isso pode causar um atraso no tempo de processamento da solicitação porque a instância tem que ser iniciada. Com o escalonamento automático, o App Engine cria automaticamente novas instâncias conforme a carga aumenta. Com o escalonamento manual, você especifica o número de instâncias para cada versão de um serviço.

O ambiente padrão do App Engine é especialmente atraente do ponto de vista de custo porque você só paga pelo que precisa e as aplicações podem escalar para zero instâncias quando não há tráfego para a aplicação.

Um serviço do App Engine recebe recursos de computação e memória baseados na classe de instância configurada para o serviço.

Para ambientes de execução de primeira geração, a classe de instância padrão para serviços front-end, chamada F1, tem até 128 MB de memória e um limite de CPU de 600 MHz. A classe de instância padrão para serviços back-end, chamada B2, tem 256 MB de memória e um limite de CPU de 1.2 GHz. Existem várias outras classes para classes de instância tanto de front-end quanto de back-end.

Para ambientes de segunda geração, F1 tem 256 MB de memória e um limite de CPU de 600 MHz, enquanto a instância B2 tem 512 MB de memória e um limite de CPU de 1.2 GHz.

Classes de instância front-end são escaladas automaticamente, e classes de instância back-end suportam escalonamento manual e básico.

Ambiente Flexível do App Engine

O ambiente flexível do App Engine oferece mais opções e controle para desenvolvedores que gostariam dos benefícios de uma plataforma como serviço (PaaS) como o App Engine, mas sem as restrições de linguagem e personalização do ambiente padrão do App Engine.

Assim como o App Engine Standard, o ambiente flexível do App Engine usa contêineres como a abstração básica do bloco de construção; no entanto, no App Engine Flexível, os usuários podem personalizar seus ambientes de execução configurando um contêiner. O ambiente flexível utiliza contêineres Docker, então desenvolvedores familiarizados com Dockerfiles podem especificar imagens base do sistema operacional, bibliotecas e ferramentas adicionais, e ferramentas personalizadas. Ele também tem suporte nativo para Java, Python, Node.js, Ruby, PHP, .NET core e Go. Veja a documentação do App Engine para versões específicas suportadas.

Em alguns aspectos, o ambiente flexível do App Engine é semelhante ao Kubernetes Engine, que será discutido na próxima seção. Ambos os produtos do Google podem executar contêineres Docker personalizados. O ambiente flexível do App Engine fornece um PaaS totalmente gerenciado e é uma boa opção quando você pode empacotar sua aplicação e serviços em um pequeno conjunto de contêineres. Esses contêineres podem então ser autoescalados de acordo com a carga. Kubernetes Engine, como você verá em breve, é projetado para gerenciar contêineres executando em um cluster que você controla. Com o Kubernetes Engine, você tem controle sobre seu cluster, mas deve monitorar e gerenciar esse cluster usando ferramentas como Cloud Monitoring e autoescalonamento. Com o ambiente flexível do App Engine, a saúde dos servidores do App Engine é monitorada pelo Google e corrigida conforme necessário sem nenhuma intervenção de sua parte.

Casos de Uso para o App Engine

O produto App Engine é uma boa escolha para uma plataforma de computação quando você tem pouca necessidade de configurar e controlar o sistema operacional subjacente ou sistema de armazenamento. O App Engine gerencia VMs e contêineres

subjacentes e alivia desenvolvedores e profissionais de DevOps de algumas tarefas comuns de administração de sistema, como aplicação de patches e monitoramento de servidores.

Quando Usar o Ambiente Padrão do App Engine

O ambiente padrão do App Engine é projetado para aplicações escritas em uma das linguagens suportadas. O ambiente padrão fornece um ambiente de execução específico para a linguagem que vem com suas próprias restrições. As restrições são menores no ambiente padrão de segunda geração do App Engine.

Se você está iniciando um novo esforço de desenvolvimento e planeja usar o ambiente padrão do App Engine, então é melhor escolher instâncias de segunda geração. Instâncias de primeira geração continuarão sendo suportadas, mas esse tipo de instância deve ser usado apenas para aplicações que já existem e foram projetadas para essa plataforma.

Quando Usar o Ambiente Flexível do App Engine

O ambiente flexível do App Engine é bem adequado para aplicações que podem ser decompostas em serviços e onde cada serviço pode ser contêinerizado. Por exemplo, um serviço pode usar uma aplicação Django para fornecer uma interface de usuário de aplicação, outro pode incorporar lógica de negócios para armazenamento de dados, e outro serviço pode agendar processamento em lote de dados carregados através da aplicação. Se você precisar instalar software adicional ou executar comandos durante a inicialização, você pode especificá-los no Dockerfile. Por exemplo, você poderia adicionar um comando run ao Dockerfile para executar apt-get update para obter a versão mais recente dos pacotes instalados.

Dockerfiles são arquivos de texto com comandos para configurar um contêiner, como especificar uma imagem base para começar e especificar comandos do gerenciador de pacotes, como apt-get e yum, para instalar pacotes.

O ambiente padrão do App Engine escala para zero instâncias em execução se não houver carga, mas isso não ocorre com o ambiente flexível. Sempre haverá pelo menos um contêiner em execução com seu serviço, e você será cobrado por esse tempo mesmo que não haja carga no sistema.

Kubernetes Engine

O Compute Engine permite que você crie e gerencie VMs individualmente ou em grupos chamados grupos de instâncias. Grupos de instâncias permitem gerenciar VMs semelhantes como uma única unidade. Isso é útil se você tiver uma frota de servidores que executam o mesmo software e têm o mesmo ciclo de vida operacional. No entanto, software moderno é frequentemente construído como uma coleção de serviços, às vezes referidos como microserviços. Serviços diferentes podem requerer diferentes configurações de VMs, mas você ainda pode querer gerenciar as várias instâncias como um único recurso, ou cluster. Você pode usar o Kubernetes Engine para isso.

Kubernetes é uma ferramenta de código aberto criada pelo Google para administrar clusters de máquinas virtuais e físicas. (Kubernetes às vezes é abreviado como K8s.) Kubernetes é um serviço de orquestração de contêineres que ajuda você a:

- Criar clusters de VMs ou máquinas físicas que executam o software de orquestração de contêineres do Kubernetes.
- Implementar aplicações contêinerizadas no cluster.
- Administrar o cluster.
- Especificar políticas, como autoescalonamento.
- Monitorar a saúde do cluster.

Kubernetes Engine é o serviço gerenciado do Kubernetes do Google Cloud. Se desejado, você poderia implantar um conjunto de VMs, instalar o Kubernetes em suas VMs e gerenciar a plataforma Kubernetes por conta própria. Com o Kubernetes Engine, você obtém os benefícios do Kubernetes sem o trabalho administrativo.

O Kubernetes Engine suporta dois modos: GKE Standard e GKE Autopilot. Com o GKE Standard, você paga por nó pelos recursos no seu cluster GKE, e você é responsável por configurar e gerenciar os nós. Com o GKE Autopilot, você paga por pod, que é uma única unidade de recursos para fornecer um serviço, e o GKE gerencia a configuração e infraestrutura.

A funcionalidade do Kubernetes é projetada para suportar clusters que executam uma variedade de aplicações. Isso difere de outras plataformas de gerenciamento de cluster que fornecem uma maneira de executar uma aplicação em vários servidores. O Spark, por exemplo, é uma plataforma de análise de big data que executa serviços Spark em um cluster de servidores. O Spark não é uma plataforma de gerenciamento de cluster de propósito geral como o Kubernetes.

O Kubernetes Engine fornece as seguintes funções:

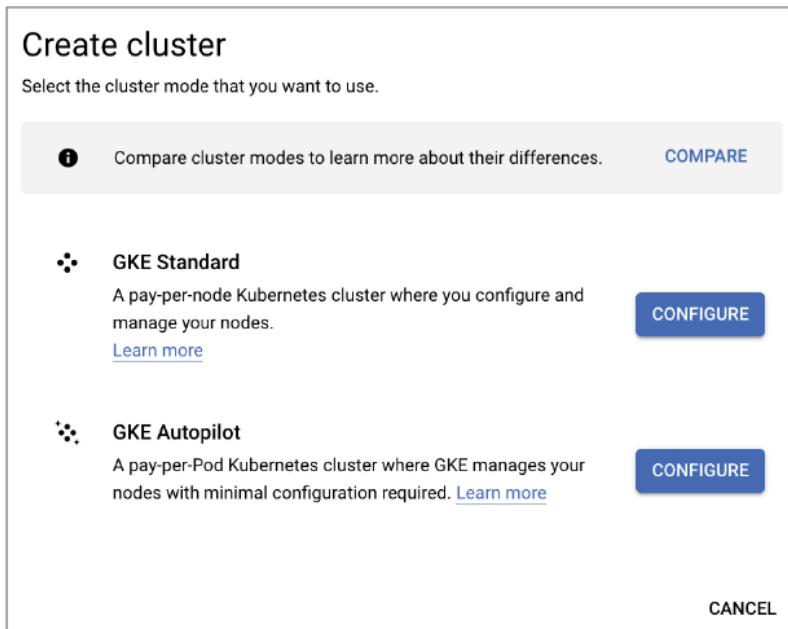
- Balanceamento de carga entre VMs do Compute Engine que são implantadas em um cluster Kubernetes
- Escalonamento automático de nós (VMs) no cluster
- Atualização automática do software do cluster conforme necessário
- Monitoramento e reparo de saúde dos nós
- Registro em log
- Suporte para pools de nós, que são coleções de nós todos com a mesma configuração

Arquitetura de Cluster Kubernetes

Um cluster Kubernetes inclui um plano de controle do cluster e um ou mais nós de trabalho. O plano de controle gerencia o cluster. Serviços do cluster, como o servidor API do Kubernetes, controladores de recursos e escalonadores, executam no plano de controle. O Servidor API do Kubernetes é o coordenador para todas as comunicações com

o cluster. O plano de controle determina quais contêineres e cargas de trabalho são executados em cada nó.

FIGURE 4.19 Kubernetes Engine supports clusters that you can manage using Standard mode, or you can have Kubernetes Engine manage many of your cluster operations using Autopilot mode.



Quando um cluster Kubernetes é criado a partir do Google Cloud Console ou de uma linha de comando, um número de nós também é criado. Estes são VMs do Compute Engine, e você pode especificar o tipo de máquina ao criar o cluster.

O Kubernetes implanta contêineres em unidades de computação abstratas conhecidas como pods. Eles geralmente têm um único contêiner, mas podem ter mais de um. Contêineres dentro de um único pod compartilham armazenamento e recursos de rede. Contêineres dentro de um pod compartilham um endereço IP e espaço de porta. Contêineres são implantados e escalados como uma unidade.

Casos de Uso do Kubernetes Engine

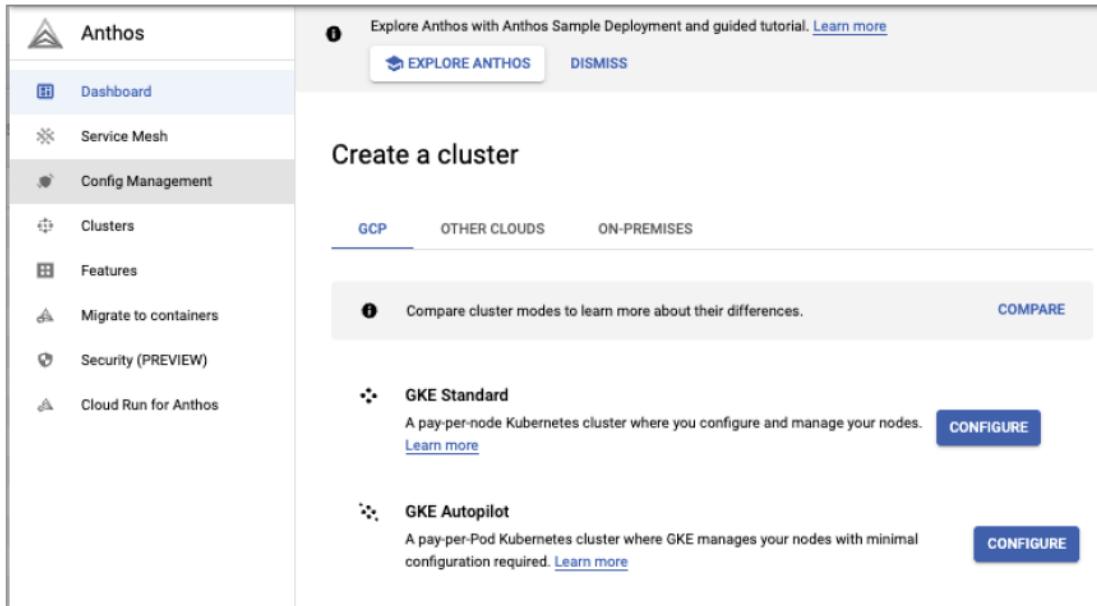
O Kubernetes Engine é uma boa escolha para aplicações em larga escala que requerem alta disponibilidade e alta confiabilidade. O Kubernetes Engine suporta o conceito de pods e conjuntos de implantação, que permitem aos desenvolvedores de aplicações e administradores gerenciar serviços como uma unidade lógica. Isso pode ajudar se você tiver um conjunto de serviços que suportam uma interface de usuário, outro conjunto que implementa lógica de negócios e um terceiro conjunto que fornece serviços de back-end. Cada um desses diferentes grupos de serviços pode ter diferentes ciclos de vida e requisitos de escalabilidade. O Kubernetes ajuda a gerenciar esses em níveis de abstração que fazem sentido para usuários, desenvolvedores e profissionais de DevOps.

Anthos

Anthos não é um serviço de computação, como o Compute Engine ou o Kubernetes Engine, mas é um serviço cada vez mais importante que é usado para

gerenciar serviços e recursos em nuvens e ambientes locais. Anthos é um serviço gerenciado para configurar e gerenciar centralmente a maneira como você implanta serviços. Com o Anthos, você pode gerenciar vários clusters GKE executando em máquinas virtuais, bem como servidores físicos. O Anthos pode gerenciar clusters executando em outras nuvens e localmente também.

FIGURE 4.20 Anthos supports the management of Kubernetes clusters in Google Cloud, other clouds, and on-premises.



Uma das vantagens de usar o Anthos é que você pode definir e aplicar políticas em diferentes ambientes. O Anthos Service Mesh é um serviço para gerenciar arquiteturas complexas de microserviços e garantir de forma consistente a segurança e o monitoramento dos serviços executados no Kubernetes.

Cloud Run

Cloud Run é um serviço gerenciado para executar contêineres. Especificamente, o Cloud Run é usado para implantar contêineres sem estado. Por sem estado, queremos dizer que qualquer instância de um contêiner executando um serviço pode responder a solicitações daquele serviço. Nenhum dado é mantido em um serviço sobre uma conexão particular ou usuário do serviço.

Cloud Run, como o App Engine, é um serviço gerenciado para executar contêineres. Quando você implanta um serviço no Cloud Run, você especifica uma imagem de contêiner, um nome de serviço, uma região, configuração de alocação de CPU, parâmetros de autoescalonamento, bem como configuração de tráfego e informações de autenticação.

Casos de Uso do Cloud Run

O ponto chave a ter em mente ao usar o Cloud Run é que o serviço executa contêineres. Isso o coloca em grupo com o Kubernetes Engine e o App Engine, que

também executam contêineres. O Cloud Run não fornece máquinas virtuais; essas são fornecidas pelo Compute Engine.

Se você está principalmente interessado em executar seu código em contêineres e não quer gerenciar a infraestrutura, então o Cloud Run é a opção recomendada se sua aplicação é sem estado.

Cloud Functions

Cloud Functions é uma plataforma de computação sem servidor projetada para executar pedaços de código de propósito único em resposta a eventos no ambiente do Google Cloud. Não há necessidade de provisionar ou gerenciar VMs, contêineres ou clusters ao usar Cloud Functions. Código escrito em Node.js, Python, Go, Java, .NET, Ruby e PHP pode ser executado no Cloud Functions. Veja a documentação do Cloud Functions para informações sobre versões suportadas dessas linguagens.

Cloud Functions não é uma plataforma de computação de propósito geral como o Compute Engine ou o App Engine. Cloud Functions fornece a "cola" entre serviços que são de outra forma independentes.

Por exemplo, um serviço pode criar um arquivo e carregá-lo no Cloud Storage, e outro serviço tem que pegar esses arquivos e realizar algum processamento no arquivo. Ambos os serviços podem ser desenvolvidos independentemente. Não há necessidade de um saber sobre o outro. No entanto, você precisará de alguma maneira de detectar que um novo arquivo foi escrito no Cloud Storage, e então a outra aplicação pode começar a processá-lo.

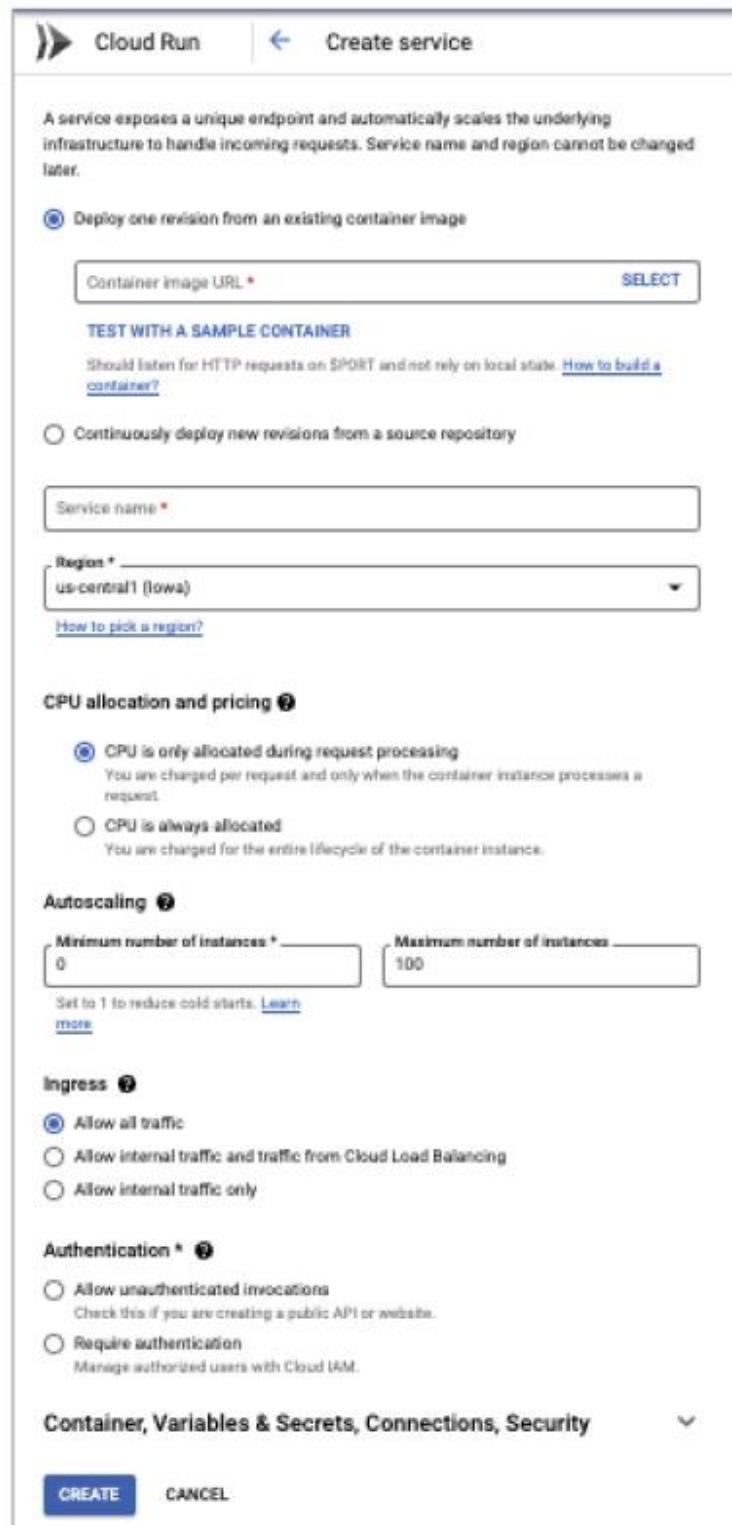
Não queremos escrever aplicações de maneiras que façam suposições sobre outros processos que podem fornecer entrada ou consumir saída. Serviços podem mudar independentemente um do outro. Não deveríamos ter que acompanhar as dependências entre serviços se pudermos evitar. Cloud Functions nos ajuda a evitar essa situação.

Ambiente de Execução do Cloud Functions

O Google Cloud gerencia tudo o que é necessário para executar seu código em um ambiente isolado e seguro. Claro, abaixo da abstração sem servidor, há servidores virtuais e físicos executando seu código, mas você, como engenheiro de nuvem, não precisa administrar nenhuma dessa infraestrutura. Três coisas-chave para lembrar sobre o Cloud Functions são as seguintes:

- As funções executam em um ambiente de execução isolado e seguro.
- Os recursos de computação escalam conforme necessário para executar tantas instâncias do Cloud Functions quanto necessário sem que você tenha que fazer nada para controlar o escalonamento.
- A execução de uma função é independente de todas as outras. Os ciclos de vida do Cloud Functions não dependem uns dos outros.

FIGURE 4.21 When deploying an application to Cloud Run, you will specify a container location to run the container, and a minimal set of configuration parameters.



Há uma corolário importante para esses pontos-chave: Cloud Functions podem estar executando em múltiplas instâncias ao mesmo tempo. Se dois usuários de aplicativos móveis carregarem um arquivo de imagem para processamento ao mesmo tempo, duas instâncias diferentes do Cloud Functions executariam aproximadamente ao

mesmo tempo. Você não precisa fazer nada para evitar conflitos entre as duas instâncias; elas são independentes.

Como cada invocação de uma função do Cloud Functions é executada em uma instância separada, as funções não compartilham memória ou variáveis. Em geral, isso significa que o Cloud Functions deve ser sem estado. Isso significa que a função não depende do estado da memória para calcular sua saída. Esta é uma restrição razoável em muitos casos, mas às vezes você pode otimizar o processamento se puder salvar o estado entre invocações. O Cloud Functions oferece algumas maneiras de fazer isso, que serão descritas no Capítulo 11, “Planejando Armazenamento na Nuvem”.

Casos de Uso do Cloud Functions

O Cloud Functions é bem adequado para processamento baseado em eventos e de curta duração. Se seus fluxos de trabalho carregam, modificam ou de outra forma alteram arquivos no Cloud Storage ou usam filas de mensagens para enviar trabalho entre serviços, então o serviço Cloud Functions é uma boa opção para executar código que inicia a próxima etapa no processamento. Algumas áreas de aplicação que se encaixam neste padrão incluem as seguintes:

- Internet das Coisas (IoT), na qual um sensor ou outro dispositivo pode enviar informações sobre o estado de um sensor. Dependendo dos valores enviados, o Cloud Functions poderia acionar um alerta ou iniciar o processamento de dados que foram carregados do sensor.
- Aplicações móveis que, como aplicativos IoT, enviam dados para a nuvem para processamento.
- Workflows assíncronos nos quais cada etapa começa algum tempo depois que a etapa anterior é concluída, mas não há suposições sobre quando as etapas de processamento serão concluídas.

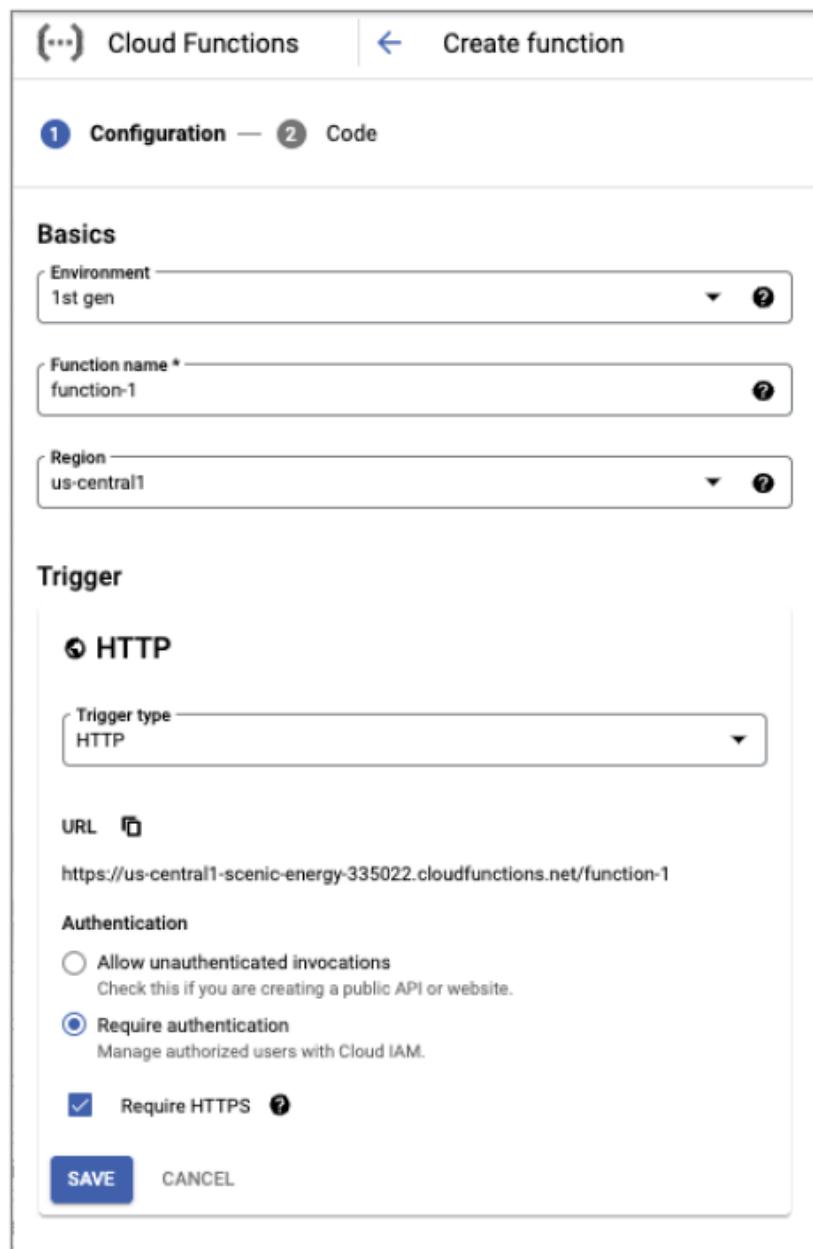
Como com outras opções de computação sem servidor, ao usar o Cloud Functions, você especifica parâmetros sobre seu serviço, neste caso uma função, e não precisa se preocupar com a infraestrutura subjacente.

Resumo

O Google Cloud oferece várias opções de computação. As opções variam no nível de controle que você, como usuário do Google Cloud, tem sobre a plataforma de computação. Geralmente, com mais controle vem mais responsabilidade e sobrecarga de gerenciamento. Seu objetivo ao escolher uma plataforma de computação é escolher uma que atenda às suas necessidades enquanto minimiza a sobrecarga de DevOps e o custo.

O Compute Engine é o serviço do Google Cloud que permite provisionar VMs. Você pode escolher entre configurações predefinidas ou pode criar uma configuração personalizada com a melhor combinação de CPUs virtuais e memória para suas necessidades. Se você pode tolerar alguma interrupção no funcionamento da VM, você pode economizar uma quantidade significativa de dinheiro usando VMs preemptíveis.

FIGURE 4.22 Configuring a Cloud Function



Aplicações de software modernas são construídas sobre múltiplos serviços que podem ter diferentes requisitos de computação e mudar em diferentes ciclos de vida. O Kubernetes Engine executa clusters de servidores que podem ser usados para executar uma variedade de serviços enquanto aloca trabalho aos servidores conforme necessário. O Kubernetes Engine também fornece monitoramento, escalonamento e remediação quando algo dá errado com uma VM no cluster.

À medida que as empresas adotam o Kubernetes e executam múltiplos clusters, elas podem recorrer ao Anthos para gerenciar clusters Kubernetes no Google Cloud, em outros clouds e localmente.

O Cloud Run é um serviço gerenciado para executar contêineres sem estado. Se você não precisa da funcionalidade completa e das características ricas do Kubernetes Engine, o Cloud Run é uma boa opção para implantar contêineres sem estado.

Aplicações fracamente acopladas podem ser unidas para implementar fluxos de trabalho complexos. Muitas vezes, queremos que cada componente seja independente dos outros. Nesses casos, frequentemente precisamos executar um código de "cola" que move a carga de trabalho de uma etapa para outra. Cloud Functions é a opção de computação sem servidor projetada para atender a essa necessidade.

Essenciais para o Exame

Entenda como as imagens são usadas para criar instâncias de VMs e como as VMs são organizadas em projetos. Instâncias executam imagens, que contêm sistemas operacionais, bibliotecas e outros códigos. Quando você cria uma instância, você especifica um projeto para conter a instância.

Saiba que o Google Cloud tem várias regiões geográficas e as regiões têm uma ou mais zonas. VMs são executadas em zonas. Uma região é uma localização geográfica, como asia-east1, europe-west2 e us-east4. As zonas dentro de uma região são ligadas por conexões de rede de alta largura de banda e baixa latência.

Entenda o que são VMs preemptíveis e quando elas são apropriadas para usar. Também entenda quando não usá-las. O Google Cloud oferece uma opção chamada VM preemptível para cargas de trabalho que podem ser interrompidas sem criar problemas.

Entenda a diferença entre os ambientes padrão e flexível do App Engine. O ambiente padrão executa uma plataforma específica de linguagem, e o ambiente flexível do App Engine permite que você execute contêineres personalizados. O App Engine é bem adequado para aplicações baseadas em HTTP(S).

Saiba que o Kubernetes é uma plataforma de orquestração de contêineres. Ele também executa contêineres em um cluster.

Entenda o Kubernetes. Ele fornece balanceamento de carga, escalonamento automático, registro em log e verificações e reparos de saúde do nó. Também saiba que o Anthos é usado para gerenciar múltiplos clusters Kubernetes através do Google Cloud, outros clouds e localmente.

Entenda o Cloud Run. O Cloud Run é um serviço gerenciado para executar contêineres sem estado e é uma boa opção quando você não precisa da funcionalidade completa do Kubernetes Engine.

Entenda o Cloud Functions. Este serviço é usado para executar programas em resposta a eventos, como o upload de um arquivo ou uma mensagem sendo adicionada a uma fila.

Questões

1. Você está implantando um aplicativo web Python no Google Cloud. O aplicativo usa apenas código personalizado e bibliotecas básicas de Python. Você espera ter um uso esporádico do aplicativo no futuro previsível e quer minimizar tanto o custo de execução do aplicativo quanto a sobrecarga de DevOps na gestão do aplicativo. Qual serviço de computação é a melhor opção para executar o aplicativo?
 - A. Compute Engine
 - B. Ambiente padrão do App Engine
 - C. Ambiente flexível do App Engine
 - D. Kubernetes Engine
2. Seu gerente está preocupado com a taxa na qual o departamento está gastando em serviços de nuvem. Você sugere que sua equipe use VMs preemptíveis para todas as seguintes opções, exceto qual?
 - A. Servidor de banco de dados
 - B. Processamento em lote sem requisito de tempo fixo para completar
 - C. Cluster de computação de alto desempenho
 - D. Nenhuma das opções acima
3. Quais parâmetros precisam ser especificados ao criar uma VM no Compute Engine?
 - A. Projeto e zona
 - B. Nome de usuário e função de admin
 - C. Conta de cobrança
 - D. Bucket do Cloud Storage
4. Sua empresa licenciou um pacote de software de terceiros que roda em Linux. Você executará várias instâncias do software em contêineres Docker. Qual dos seguintes serviços do Google Cloud você poderia usar para implantar este pacote de software?
 - A. Apenas Compute Engine
 - B. Apenas Kubernetes Engine
 - C. Apenas Compute Engine, Kubernetes Engine e o ambiente flexível do App Engine
 - D. Compute Engine, Kubernetes Engine, o ambiente flexível do App Engine ou o ambiente padrão do App Engine

5. Você pode especificar pacotes para instalar em um contêiner Docker incluindo comandos em qual arquivo?
 - A. Docker.cfg
 - B. Dockerfile
 - C. Config.dck
 - D. install.cfg
6. Qual dos seguintes pode ser gerenciado usando o Anthos?
 - A. Clusters Kubernetes apenas no Google Cloud
 - B. Contêineres do App Engine Flexível e clusters Kubernetes no Google Cloud
 - C. Contêineres do App Engine Flexível, Cloud Functions e clusters Kubernetes no Google Cloud
 - D. Clusters Kubernetes no Google Cloud, AWS e localmente
7. Seu gerente está fazendo uma apresentação para os executivos da sua empresa defendendo que vocês começem a usar o Kubernetes Engine. Você sugere que o gerente destaque todos os recursos que o Kubernetes fornece para reduzir a carga de trabalho dos engenheiros de DevOps. Você descreve vários recursos, incluindo todos os seguintes, exceto qual?
 - A. Balanceamento de carga entre as VMs do Compute Engine que são implantadas em um cluster Kubernetes
 - B. Varredura de segurança para vulnerabilidades
 - C. Escalonamento automático de nós no cluster
 - D. Atualização automática do software do cluster conforme necessário
8. Sua empresa está prestes a lançar um serviço online que se baseia em uma nova experiência de interface de usuário impulsionada por um conjunto de serviços que serão executados em seus servidores. Um conjunto separado de serviços gerencia a autenticação e autorização. Um terceiro conjunto de serviços mantém o controle das informações da conta. Todos os três conjuntos de serviços devem ser altamente confiáveis e escaláveis para atender à demanda. Qual dos serviços do Google Cloud é a melhor opção para implantar isso?
 - A. Ambiente padrão do App Engine
 - B. Compute Engine
 - C. Cloud Functions
 - D. Kubernetes Engine

9. Um aplicativo móvel faz upload de imagens para análise, incluindo a identificação de objetos na imagem e a extração de texto que pode estar embutido na imagem. Um terceiro criou o aplicativo móvel, e você desenvolveu o serviço de análise de imagem. Ambos concordam em usar o Cloud Storage para armazenar imagens. Você quer manter os dois serviços completamente desacoplados, mas precisa de uma maneira de invocar a análise de imagem assim que uma imagem for carregada. Como isso deve ser feito?
- A. Alterar o aplicativo móvel para iniciar uma VM executando o serviço de análise de imagem e fazer com que essa VM copie o arquivo do armazenamento para o armazenamento local na VM. Ter o serviço de imagem executado na VM.
- B. Escrever uma função em Python que é invocada pelo Cloud Functions quando um novo arquivo de imagem é escrito no bucket do Cloud Storage que recebe novas imagens. A função deve enviar a URL do arquivo carregado para o serviço de análise de imagem. O serviço de análise de imagem então carregará a imagem do Cloud Storage, realizará a análise e gerará resultados, que podem ser salvos no Cloud Storage.
- C. Ter um cluster Kubernetes rodando continuamente, com um pod dedicado à listagem do conteúdo do bucket de upload e à detecção de novos arquivos no Cloud Storage e outro pod dedicado à execução do software de análise de imagem.
- D. Ter uma VM do Compute Engine rodando e listando continuamente o conteúdo do bucket de upload no Cloud Storage para detectar novos arquivos. Outra VM deve estar continuamente executando o software de análise de imagem.
10. Sua equipe está desenvolvendo um novo pipeline para analisar um fluxo de dados de sensores em dispositivos de fabricação. O pipeline antigo ocasionalmente corrompia dados porque threads paralelos sobrescreviam dados escritos por outros threads. Você decide usar o Cloud Functions como parte do pipeline. Como desenvolvedor de uma Cloud Function, o que você deve fazer para evitar que múltiplas invocações da função interfiram umas com as outras?
- A. Incluir uma verificação no código para garantir que outra invocação não esteja sendo executada ao mesmo tempo.
- B. Agendar cada invocação para ser executada em um processo separado.
- C. Agendar cada invocação para ser executada em uma thread separada.
- D. Nada. O Google Cloud garante que as invocações de funções não interfiram umas com as outras.
11. Um cliente seu processa informações pessoais e de saúde para hospitais. Todas as informações de saúde precisam ser protegidas de acordo com regulamentos governamentais. Seu cliente quer mover sua aplicação para o Google Cloud, mas quer usar a biblioteca de criptografia que eles usaram no passado. Você sugere que todas as VMs executando a aplicação tenham a biblioteca de criptografia instalada. Que tipo de imagem você usaria para isso?
- A. Imagem personalizada

B. Imagem pública

C. CentOS 6 ou 7

D. Ubuntu 18 ou posterior

12. Qual é o nível mais baixo da hierarquia de recursos?

A. Pasta

B. Projeto

C. Arquivo

D. Instância de VM

13. Sua empresa está vendo um aumento marcante na taxa de crescimento de clientes na Europa. A latência está se tornando um problema porque sua aplicação está rodando em us-central1. Você sugere implantar seus serviços em uma região na Europa. Você tem várias opções. Você deve considerar todos os seguintes fatores, exceto qual?

A. Custo

B. Latência

C. Regulamentações

D. Confiabilidade

14. Qual papel dá aos usuários controle total sobre instâncias do Compute Engine?

A. Papel de Gerente de Computação

B. Papel de Admin de Computação

C. Papel de Gerente Regional de Computação

D. Papel de Admin de Segurança de Computação

15. Quais das seguintes são limitações de uma VM preemptível?

A. Será encerrada dentro de 24 horas.

B. Pode nem sempre estar disponível. A disponibilidade pode variar entre zonas e regiões.

C. Não pode migrar para uma VM regular.

D. Todas as opções acima.

16. Qual dos seguintes eliminaria o Cloud Run como uma opção para implantar uma aplicação no Google Cloud?

A. A aplicação usa uma mistura de código de aplicação Java e Python.

- B. A aplicação armazena dados sobre uma sessão na memória para uso em várias solicitações durante uma sessão.
- C. A aplicação roda em um contêiner.
- D. A configuração do contêiner é especificada em um Dockerfile.
17. Ao usar o ambiente padrão do App Engine, qual dos seguintes runtimes de linguagem não é suportado?
- A. Java
- B. Python
- C. C
- D. Go
18. Você quer garantir que todos os serviços executando em um cluster do Kubernetes Engine usem os mesmos serviços de autenticação e monitoramento. Qual serviço você usaria?
- A. Cloud Functions
- B. Anthos Service Mesh
- C. App Engine Flexível
- D. App Engine Padrão
19. Você está implantando um conjunto de máquinas virtuais no Compute Engine. Você quer garantir que malware não comprometa o sistema operacional, então você quer validar a integridade da inicialização. Qual recurso do Compute Engine você ativaria?
- A. Chaves de criptografia fornecidas pelo cliente
- B. vTPM
- C. Tenência exclusiva
- D. Papéis de gerenciamento de identidade e acesso
20. Um cliente o contratou para ajudar a reduzir sua sobrecarga de DevOps. Engenheiros estão gastando muito tempo aplicando patches em servidores e otimizando a utilização do servidor. Eles querem migrar para plataformas sem servidor tanto quanto possível. Seu cliente ouviu falar do Cloud Functions e quer usá-los. Você recomendaria todos os seguintes tipos de aplicações, exceto qual?
- A. Procedimentos de carga de dados de data warehouse de longa duração
- B. Processamento de back-end de IoT
- C. Processamento de eventos de aplicativos móveis
- D. Workflows assíncronos

CAPÍTULO 5

Computação com o Compute Engine Máquinas Virtuais

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

✓✓ 1.3 Instalar e configurar a interface de linha de comando (CLI), especificamente o Cloud SDK (por exemplo, definindo o projeto padrão)

✓✓ 2.2 Planejar e configurar recursos de computação. Considerações incluem:

■■ Selecionar opções de computação apropriadas para uma determinada carga de trabalho (por exemplo, Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)

■■ Usar VMs preemptíveis e tipos de máquina personalizados conforme apropriado

Neste capítulo, você aprenderá sobre o Google Cloud Console, uma interface gráfica de usuário para trabalhar com o Google Cloud. Você aprenderá como instalar o Google Cloud SDK e usá-lo para criar instâncias de máquinas virtuais e como usar o Cloud Shell como uma alternativa para instalar o Google Cloud SDK localmente.

Criando e Configurando Máquinas Virtuais com o Console

Vamos criar uma VM no Compute Engine. Temos três opções para fazer isso: podemos usar o Google Cloud Console, o Google Cloud Software Development Kit (SDK) ou o Google Cloud Shell. Vamos começar com o console.

O Google Cloud Console é uma interface gráfica de usuário (GUI) baseada na web para criar, configurar e gerenciar recursos no Google Cloud. Neste capítulo, usaremos para criar uma VM.

Para abrir o console, navegue no seu navegador até <https://console.cloud.google.com> e faça login. A Figura 5.1 mostra um exemplo do formulário principal no console.

No console, selecione a opção Selecionar Um Projeto para exibir os projetos existentes. Você também pode criar um novo projeto a partir deste formulário, conforme mostrado na Figura 5.2.

FIGURE 5.1 The main starting form of Google Cloud Console

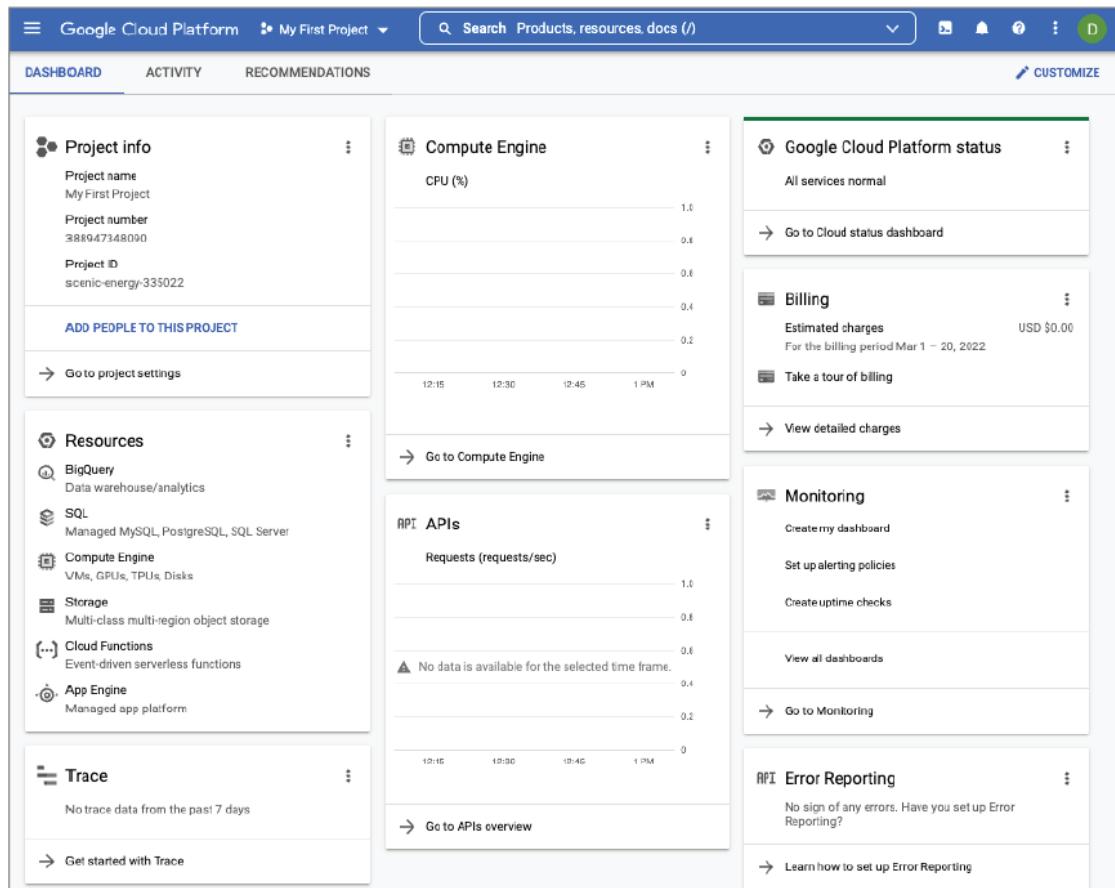


FIGURE 5.2 The Project form lets you choose the project you want to work with when creating VMs. You can also create a new project here.

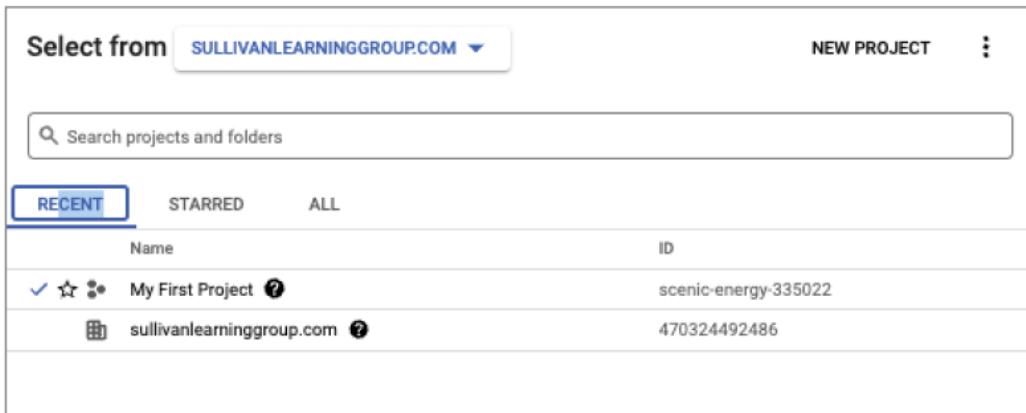
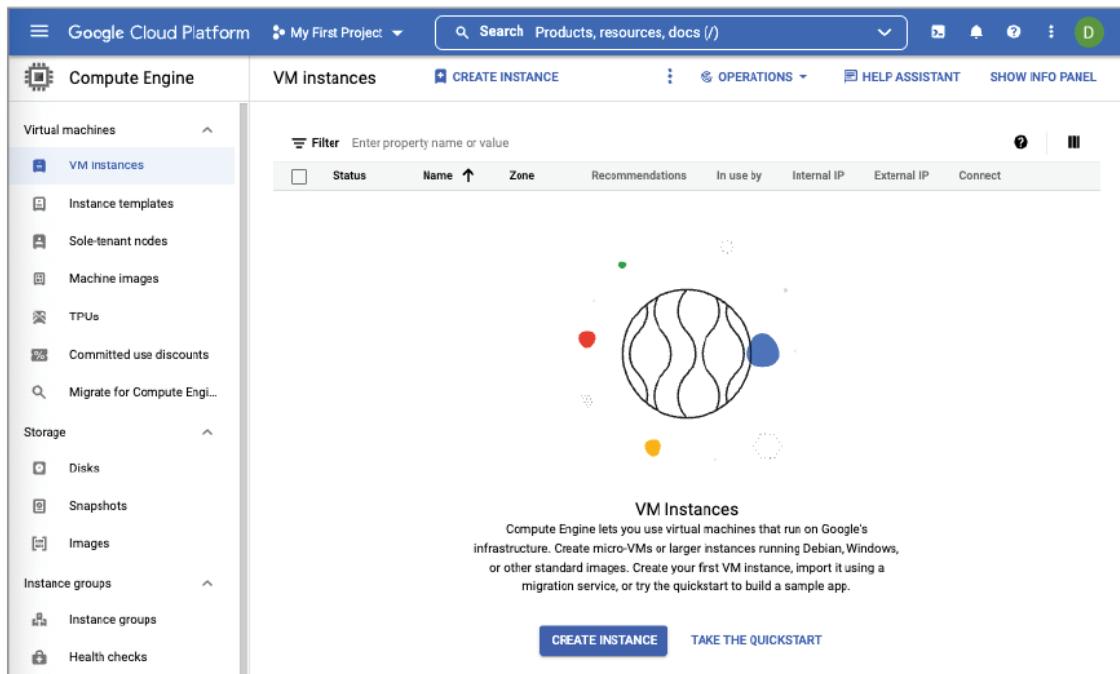


FIGURE 5.3 The starting panel for creating a VM



Principais Detalhes da Configuração da Máquina Virtual

Dentro do console, você pode especificar todos os detalhes necessários sobre a configuração da VM que você está criando, incluindo os seguintes:

- Nome da VM
- Região e zona onde a VM será executada
- Tipo de máquina, que determina o número de CPUs e a quantidade de memória na VM
- Disco de inicialização, que inclui o sistema operacional que a VM executará

Você pode escolher o nome da sua VM. Isso é principalmente para o seu uso. O Google Cloud usa outros identificadores internamente para gerenciar as VMs.

Você precisará especificar uma região. Regiões são grandes áreas geográficas. Uma lista parcial de regiões é mostrada na Figura 5.5.

Depois de selecionar uma região, você pode selecionar uma zona. Lembre-se, uma zona é uma instalação semelhante a um centro de dados dentro de uma região. A Figura 5.6 mostra um exemplo de lista de zonas disponíveis na região us-east-1.

FIGURE 5.4 Part of the main configuration form for creating VMs in Compute Engine

The screenshot shows the 'Create VM' configuration page. At the top, there's a 'Name*' field containing 'instance-1', a 'Labels' section with a '+ ADD LABELS' button, and a 'Monthly estimate' section showing '\$25.46' (That's about \$0.03 hourly) with a note about pay-as-you-go billing. Below these are 'Region*' and 'Zone*' dropdowns set to 'us-central1 (Iowa)' and 'us-central1-a' respectively, with notes that they are permanent. A 'DETAILS' section is partially visible. The main configuration area is titled 'Machine configuration' and includes a 'Machine family' tab bar with 'GENERAL-PURPOSE' selected (the others are COMPUTE-OPTIMIZED, MEMORY-OPTIMIZED, and GPU). Under 'GENERAL-PURPOSE', it says 'Machine types for common workloads, optimized for cost and flexibility'. A 'Series' dropdown is set to 'E2'. A note says 'CPU platform selection based on availability'. A 'Machine type' dropdown is set to 'e2-medium (2 vCPU, 4 GB memory)'. Below this, there's a summary table for the e2-medium type: **vCPU** (1 shared core), **Memory** (4 GB), and a small icon of stacked CPU cores. A 'CPU PLATFORM AND GPU' section is collapsed. The 'Display device' section is collapsed, with a note about enabling screen capturing and recording tools and an unchecked checkbox for 'Enable display device'. The 'Confidential VM service' section is collapsed, with a note about enabling the Confidential Computing service and an unchecked checkbox. The 'Container' section is collapsed, with a note about deploying a container image and a 'DEPLOY CONTAINER' button. The entire form has rounded corners and a light gray background.

FIGURE 5.5 A partial list of regions providing Compute Engine services

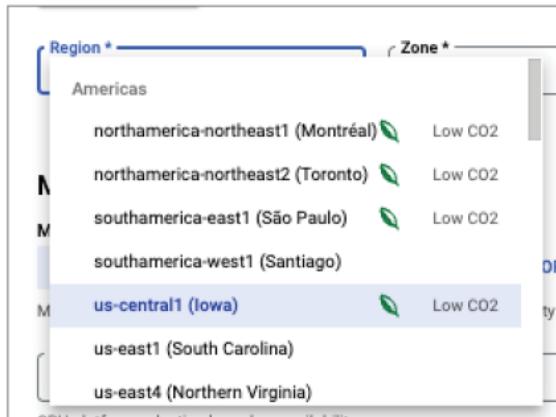
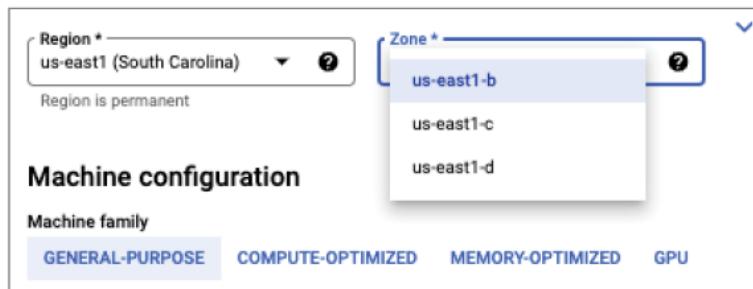


FIGURE 5.6 A list of zones within the us-east1 region



Depois de especificar uma região e uma zona, o Google Cloud pode determinar as VMs disponíveis nessa zona. Nem todas as zonas têm a mesma disponibilidade. A Figura 5.7 mostra um exemplo da listagem de tipos de máquina disponíveis para a série E2 na zona us-east1-b.

O Google Cloud organiza as máquinas virtuais em famílias de máquinas, séries e tipos de máquina. Uma família de máquinas é um conjunto de configurações de processador e hardware projetadas para cargas de trabalho específicas, como uso geral, otimizado para computação e otimizado para memória. Dentro de uma família, as máquinas são organizadas em séries e gerações, conforme mostrado na Figura 5.8.

Dentro de uma série, você terá a opção de um ou mais tipos de máquina, que variam com base no número de CPUs virtuais e na quantidade de memória.

Para aplicações e serviços que exigem alta segurança, você pode habilitar o Serviço de VM Confidencial, que mantém os dados na memória criptografados usando chaves de criptografia às quais o Google não tem acesso.

FIGURE 5.7 A partial list of machine types available in the us-east1-b zone

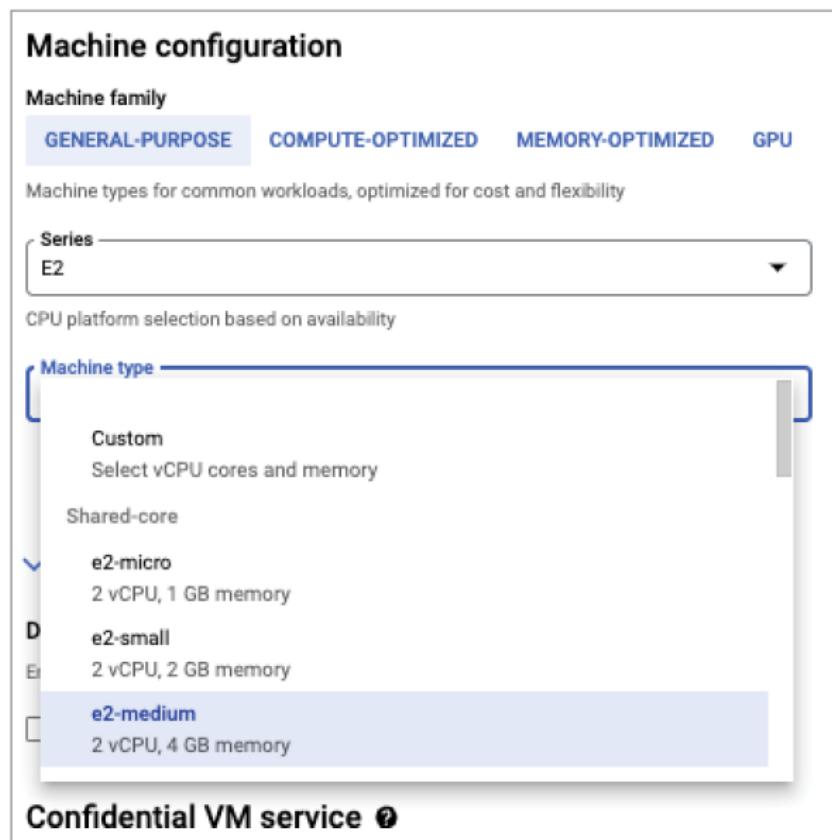
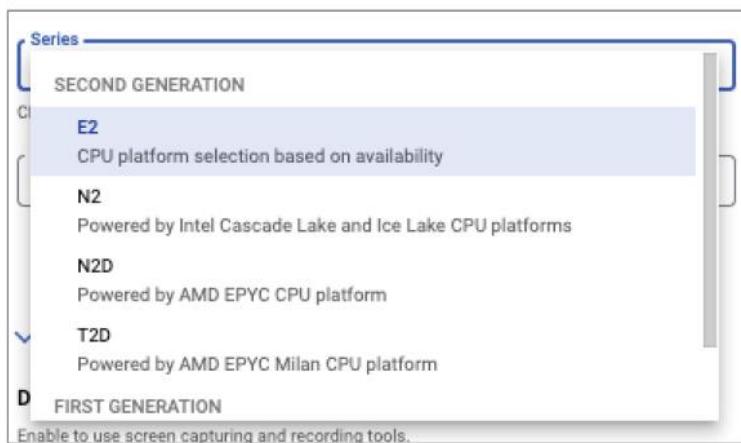


FIGURE 5.8 Virtual machines within a machine family are further organized into series and generations based on the type of processor.



Você tem a opção de escolher executar um contêiner em sua máquina virtual. Se decidir fazer isso, você deve especificar um contêiner que esteja em um repositório público ou no Google Container Registry. Isso pode ser útil se você quiser executar um contêiner com software especializado ou alguma configuração personalizada.

A seção Disco de Inicialização lista uma configuração padrão. Clicar no botão Alterar abre o formulário do Disco de Inicialização, conforme mostrado na Figura 5.9.

FIGURE 5.9 Form for configuring the boot disk of the VM

The screenshot shows a 'Boot disk' configuration interface. At the top, there's a message: 'Select an image or snapshot to create a boot disk; or attach an existing disk. Can't find what you're looking for? Explore hundreds of VM solutions in [Marketplace](#)'. Below this are four tabs: 'PUBLIC IMAGES' (which is selected), 'CUSTOM IMAGES', 'SNAPSHOTS', and 'EXISTING DISKS'. The main area contains the following fields:

- 'Operating system': A dropdown menu showing 'Debian'.
- 'Version *': A dropdown menu showing 'Debian GNU/Linux 10 (buster)'. Below it, a note says 'amd64 built on 20220317, supports Shielded VM features'.
- 'Boot disk type *': A dropdown menu showing 'Balanced persistent disk'.
- 'Size (GB) *': An input field containing '10'.

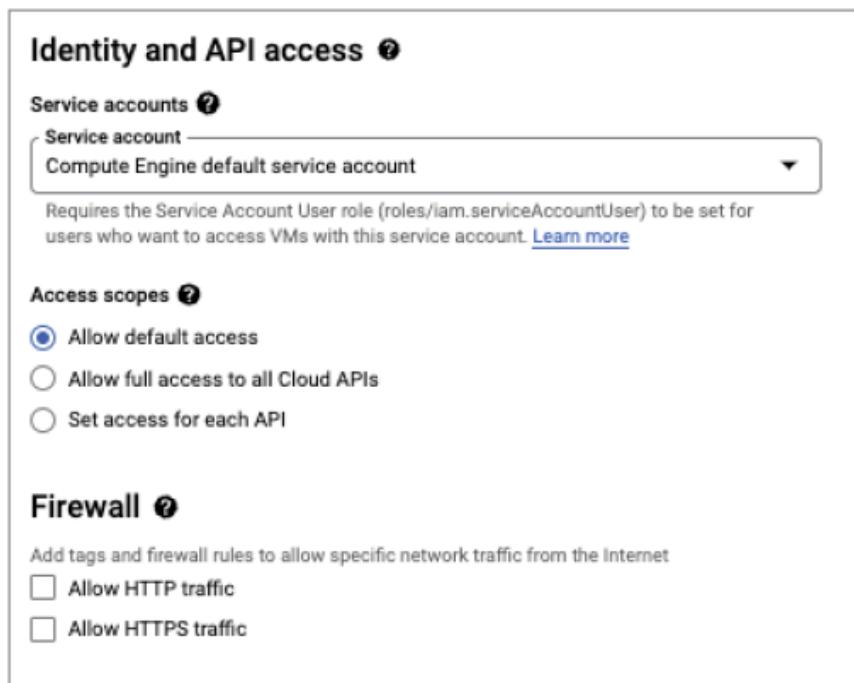
At the bottom left is a 'SHOW ADVANCED CONFIGURATION' link, and at the bottom right are 'SELECT' and 'CANCEL' buttons.

Aqui você pode escolher o sistema operacional que deseja usar. Você também pode escolher o tipo de disco de inicialização: Disco Persistente Equilibrado, Disco Persistente Extremo, Disco Persistente SSD ou Disco Persistente Padrão. Você também pode especificar o tamanho do disco.

- Discos persistentes equilibrados usam unidades de estado sólido (SSDs) e equilibram custo e desempenho.
- Discos persistentes extremos usam SSDs, mas oferecem alto desempenho e permitem que você provisione o nível desejado de operações de entrada e saída por segundo (IOPS).
- Discos persistentes SSD usam unidades de estado sólido.
- Discos persistentes padrão usam discos rígidos padrão (HDDs).

Após a seção de Disco de Inicialização está a seção Identidade e Acesso à API (veja Figura 5.10). Aqui você pode especificar uma conta de serviço para a VM e definir o escopo de acesso à API. Se você quiser que os processos em execução nesta VM usem apenas algumas APIs, você pode usar essas opções para limitar o acesso da VM a APIs específicas.

FIGURE 5.10 Identity And API Access and Firewall configurations



Na próxima seção, você pode selecionar se deseja que a VM aceite tráfego HTTP ou HTTPS.

Detalhes da Configuração Avançada

Clique em Gerenciamento, Segurança, Discos, Rede e Tenência Única para exibir opções avançadas de configuração. Isso expandirá uma lista de opções avançadas de configuração.

Aba de Gerenciamento

A aba de Gerenciamento do formulário (Figura 5.11) fornece um espaço onde você pode descrever a VM e seu uso. Você também pode criar etiquetas, que são pares chave-valor. Você pode atribuir qualquer etiqueta que desejar. Etiquetas e uma descrição geral são frequentemente usadas para ajudar a gerenciar suas VMs e ilustrar como elas estão sendo usadas. As etiquetas são particularmente importantes quando o número de seus servidores cresce. É uma prática recomendada incluir uma descrição e etiquetas para todas as VMs.

FIGURE 5.11 The first part of the Management tab of the VM creation form

Management
Description, deletion protection, reservations, automation, and availability policies

Description

Deletion protection ?
 Enable deletion protection

Reservation name
Automatically use created reservation ▾
Use an existing reservation when creating this VM instance

Automation

Startup script

You can choose to specify a startup script that will run when your instance boots up or restarts. Startup scripts can be used to install software and updates, and to ensure that services are running within the virtual machine. [Learn more](#)

Metadata
You can set custom metadata for an instance or project outside of the server-defined metadata. This is useful for passing in arbitrary values to your project or instance that can be queried by your code on the instance. [Learn more](#)

+ ADD ITEM

Availability policy

Preemptibility
Off (Recommended) ▾
A preemptible VM costs much less, but lasts only 24 hours. It can be terminated sooner due to system demands. [Learn more](#)

On host maintenance
Migrate VM instance (Recommended) ▾
When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart
On (recommended) ▾
Compute Engine can automatically restart VM instances if they are terminated for non-user-initiated reasons (maintenance event, hardware failure, software failure and so on)

Se você quiser forçar uma confirmação extra antes de excluir uma instância, você pode selecionar a opção de Proteção contra Exclusão. Se alguém tentar excluir a instância, a operação falhará.

Se você tiver reservado recursos de instância do Compute Engine, eles serão automaticamente usados, mas você pode indicar que reservas não devem ser usadas para uma instância particular.

Você pode especificar um script de inicialização para ser executado quando a instância iniciar. Copie o conteúdo do script de inicialização para a caixa de texto de Automação. Por exemplo, você poderia colar um script Bash ou Python diretamente na caixa de texto.

A seção de Metadados permite que você especifique pares chave-valor associados à instância. Esses valores são armazenados em um servidor de metadados, que está disponível para consulta usando a API do Compute Engine. As tags de metadados são especialmente úteis se você tem um script comum que deseja executar na inicialização ou desligamento, mas quer que o comportamento do script varie de acordo com alguns valores de metadados.

Em Política de Disponibilidade, existem três menus suspenso:

- Modelo de Provisionamento de VM, que pode ser padrão ou spot. Spot permite que o Google desligue o servidor com um aviso de 30 segundos. Em troca, o custo de um servidor preemptível é muito menor do que o de um servidor não preemptível.
- Manutenção no Host, que indica se o servidor virtual deve ser migrado para outro servidor físico quando ocorrer um evento de manutenção.
- Reinício Automático, que indica se o servidor deve ser reiniciado automaticamente se parar por causa de uma falha de hardware, evento de manutenção ou algum outro evento não controlado pelo usuário.

Aba de Segurança

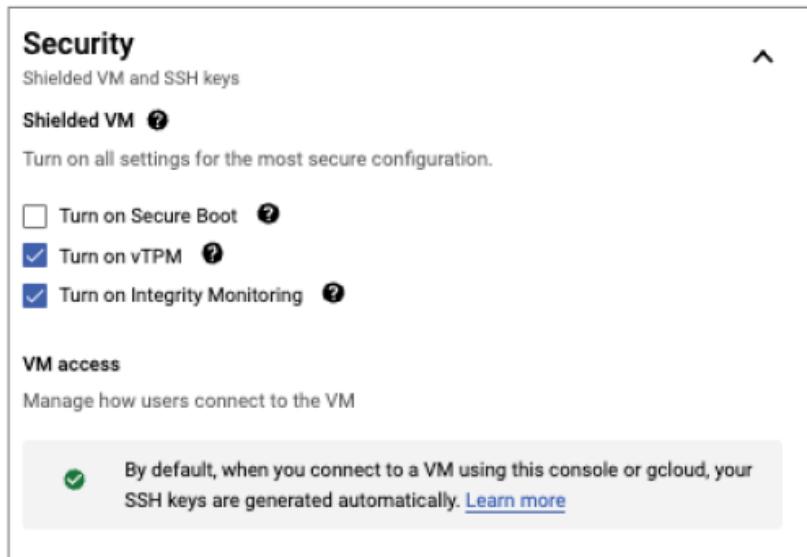
Na aba de Segurança, você pode especificar se deseja usar VMs Blindadas (Shielded VMs) e chaves de Shell Seguro (SSH).

As VMs Blindadas são configuradas para ter mecanismos de segurança adicionais que você pode escolher executar (veja Figura 5.12). Isso inclui o seguinte:

- Inicialização Segura, que garante que apenas software de sistema operacional autenticado seja executado na VM. Isso é feito verificando as assinaturas digitais do software. Se uma verificação de assinatura falhar, o processo de inicialização será interrompido.
- Módulo de Plataforma Confiável Virtual (vTPM), que é uma versão virtualizada de um Módulo de Plataforma Confiável (TPM). Um TPM é um chip de computador especializado projetado para proteger recursos de segurança, como chaves e certificados.
- Monitoramento de Integridade, que usa uma linha de base conhecida de medições de inicialização para comparar com medições de inicialização recentes. Se a verificação falhar, isso significa que existe alguma diferença entre a medição de linha de base e as medições atuais.

O Google Cloud suporta o conceito de chaves SSH em todo o projeto, que são usadas para dar aos usuários acesso em todo o projeto às VMs. Você pode bloquear esse comportamento na VM se usar chaves SSH de todo o projeto e não quiser que todos os usuários do projeto tenham acesso a esta máquina.

FIGURE 5.12 You can place additional security controls on VMs.



Discos de Inicialização e Discos Adicionais

Na aba de Disco de Inicialização da página Criar Instância, você pode especificar opções de configuração avançadas, conforme mostrado na Figura 5.13. Em Regra de Exclusão, você pode especificar se o disco de inicialização deve ser excluído quando a instância for excluída. Você também pode selecionar como gostaria de gerenciar as chaves de criptografia para o disco de inicialização. Por padrão, o Google gerencia essas chaves.

Na aba de configuração de disco, você também tem a opção de adicionar um novo disco ou anexar um disco existente. A Figura 5.14 mostra a aba para adicionar um novo disco.

Ao adicionar um novo disco, o formulário na Figura 5.14 aparece. Aqui, você especifica um nome e descrição e informações de origem. A origem especifica se você deseja usar um disco em branco ou criar um usando um snapshot ou imagem. Você também especifica o tamanho e o tipo do disco. Se você deseja fazer backup automático do seu disco, pode especificar um cronograma de snapshot. Por padrão, o Google gerenciará as chaves de criptografia para o disco, mas você também pode especificar chaves de criptografia gerenciadas pelo cliente (CMEKs) ou chaves de criptografia fornecidas pelo cliente (CSEKs).

Adicionar um disco existente exibe o formulário mostrado na Figura 5.15. Aqui você escolhe um disco de uma lista de discos existentes e especifica se o disco deve ser anexado como Leitura/Escrita ou Somente Leitura.

Você também pode especificar se o disco deve ser excluído quando a VM for excluída. O padrão é manter o disco. Por fim, você pode fornecer um nome personalizado para o disco.

FIGURE 5.13 Boot disk advanced configuration

Boot disk

Select an image or snapshot to create a boot disk; or attach an existing disk. Can't find what you're looking for? Explore hundreds of VM solutions in [Marketplace](#)

PUBLIC IMAGES **CUSTOM IMAGES** **SNAPSHOTS** **EXISTING DISKS**

Operating system — Debian

Version * — Debian GNU/Linux 10 (buster)

amd64 built on 20220317, supports Shielded VM features

Boot disk type * — Balanced persistent disk **Size (GB) *** — 10

Deletion rule
When deleting instance
 Keep boot disk
 Delete boot disk

Encryption
Data is encrypted automatically. Select an encryption key management solution.
 Google-managed encryption key
No configuration required
 Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service
 Customer-supplied encryption key (CSEK)
Manage outside of Google Cloud

Snapshot schedule
Use snapshot schedules to automate disk backups. [Learn more](#)

Select a snapshot schedule

Device name ?
Used to reference the device for mounting or resizing.
 Use a custom device name
Device name — instance-1
Based on instance name (default)

[▲ HIDE ADVANCED CONFIGURATION](#)

FIGURE 5.14 Adding a new disk to a Compute Engine instance

Add new disk X

Name * ?
Name is permanent

Description

Source
Create a blank disk, apply a bootable disk image, or restore a snapshot of another disk in this project.

Disk source type * ▼ ?

Disk settings

Disk type * ▼ ?

[COMPARE DISK TYPES](#)

Size * GB ?
Provision between 10 and 65,536 GB

Snapshot schedule (Recommended)
Use snapshot schedules to automate disk backups. [Learn more](#)

Select a snapshot schedule ▼

Encryption
Data is encrypted automatically. Select an encryption key management solution.

Google-managed encryption key
No configuration required

Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

Customer-supplied encryption key (CSEK)
Manage outside of Google Cloud

Labels ?

[+ ADD LABEL](#)

SAVE CANCEL

FIGURE 5.15 Form for adding an existing disk to a VM

The screenshot shows a configuration dialog titled "Existing disk". At the top, there is a dropdown menu labeled "Disk *". Below it, the section "Attachment settings" contains two groups of options: "Mode" (with "Read/write" selected) and "Deletion rule" (with "Keep disk" selected). Under "Device name", there is a note about referencing the device for mounting or resizing, a checked checkbox for "Use a custom device name", and a text input field containing "persistent-disk-1". At the bottom of the dialog are "SAVE" and "CANCEL" buttons.

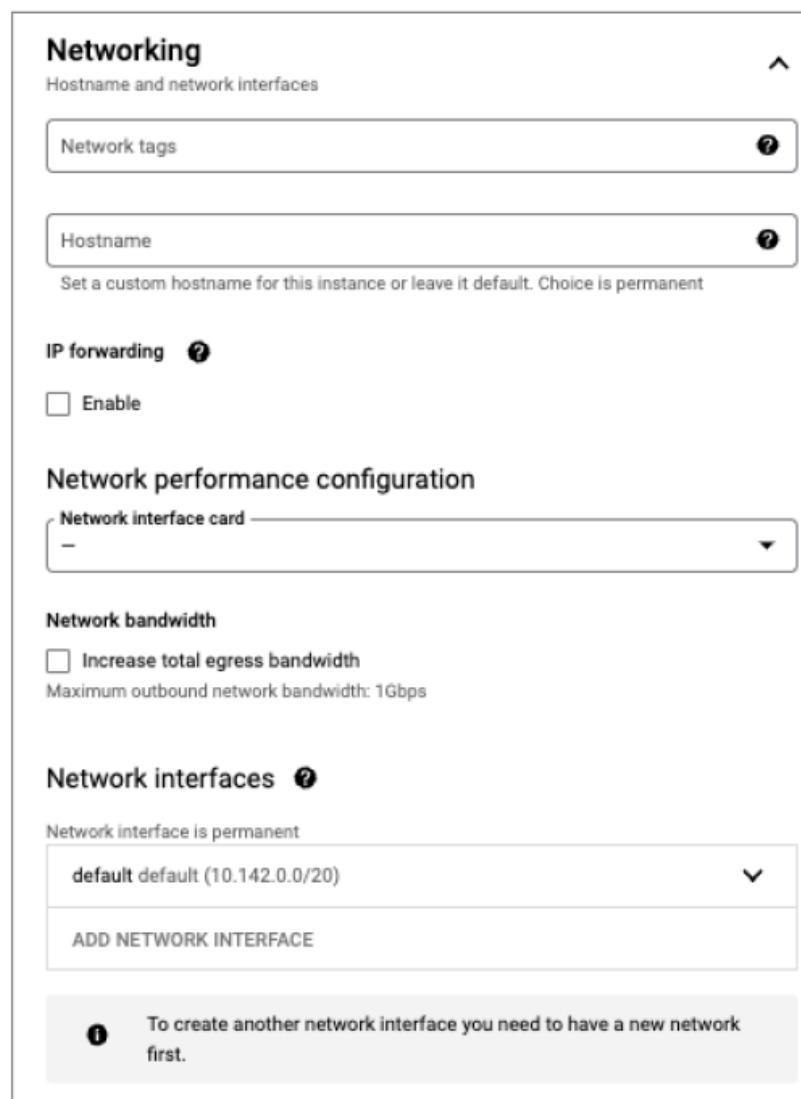
Aba de Rede

Na aba de Rede, você pode ver as informações da interface de rede, incluindo o endereço IP da VM. Se você tem duas redes, você tem a opção de adicionar outra interface de rede àquela outra rede. O uso de interfaces de rede duplas pode ser útil se você estiver executando algum tipo de proxy ou servidor que atua como um controle para o fluxo de algum tráfego entre as redes. Além disso, você pode adicionar tags de rede nesta aba (veja a Figura 5.16).

Tenência Exclusiva

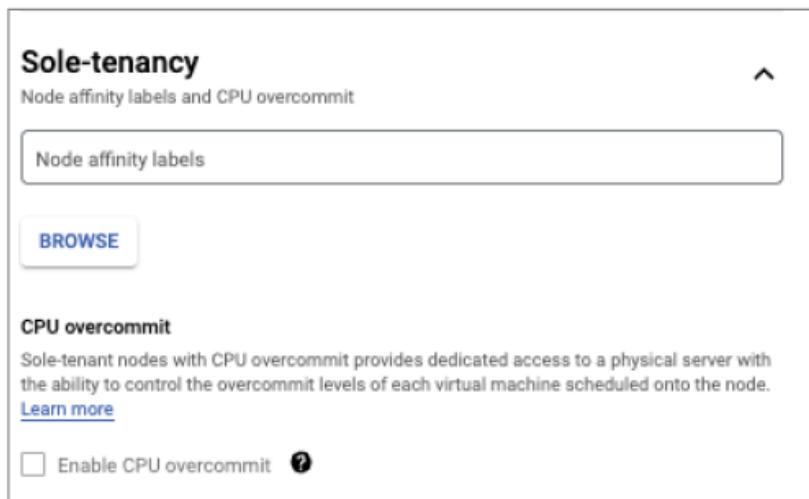
Se você precisa garantir que suas VMs rodem em um servidor apenas com suas outras VMs, então você pode especificar tenência exclusiva. A aba de Tenência Exclusiva permite que você especifique etiquetas sobre tenência exclusiva para o servidor (veja a Figura 5.17).

FIGURE 5.16 Options for network configuration of a VM



Nós de Tenência Única podem ser configurados para permitir o superdimensionamento de recursos de CPU, mas isso só é permitido em máquinas com quatro ou mais CPUs. Etiquetas de afinidade de nó são usadas para determinar onde uma VM pode ser executada.

FIGURE 5.17 Sole tenancy configuration options



Criando e Configurando Máquinas Virtuais com o Cloud SDK

Uma segunda maneira de criar e configurar VMs é com o Google Cloud SDK, que fornece uma interface de linha de comando (CLI). Para usar o Cloud SDK, primeiro você precisará instalá-lo em seu dispositivo local.

Instalando o Cloud SDK

Você tem três opções para interagir com recursos do Google Cloud:

- Usando uma interface de linha de comando
- Usando uma interface RESTful
- Usando o Cloud Shell

Antes de usar as duas primeiras opções do seu sistema local, você precisará instalar o Cloud SDK em sua máquina. O Cloud Console é uma GUI que você pode acessar através de um navegador em <https://console.cloud.google.com>.

O Cloud SDK pode ser instalado em computadores Linux, Windows ou Mac.

Instalando o Cloud SDK no Linux

Se você estiver usando Linux, pode instalar o Cloud SDK usando o gerenciador de pacotes do seu sistema operacional. Ubuntu e outras distribuições Debian usam apt-get para instalar pacotes. Red Hat Enterprise, CentOS e outras distribuições Linux usam yum. Para instruções sobre como usar apt-get, veja <https://cloud.google.com/sdk/docs/install-sdk#deb>. Para instruções sobre a instalação no Red Hat Enterprise ou CentOS, veja <https://cloud.google.com/sdk/docs/install-sdk#rpm>.

Cloud SDK no macOS

Instruções para instalar no Mac e o arquivo de instalação para o Cloud SDK estão disponíveis em <https://cloud.google.com/sdk/docs/install-sdk#mac>. O primeiro passo é verificar se você tem o Python 3 instalado. Existem três versões do Cloud SDK, uma para

macOS de 32 bits; uma para macOS de 64 bits rodando em processadores x86; e uma para macOS de 64 bits rodando em arm64, o processador Apple M1.

Instalando o Cloud SDK no Windows

Para instalar o Cloud SDK em uma plataforma Windows, você precisará baixar o instalador apropriado. Você pode encontrar instruções em <https://cloud.google.com/sdk/docs/install-sdk#windows>.

Exemplo de Instalação no Ubuntu Linux

O primeiro passo para instalar o Cloud SDK é obter a versão apropriada do pacote para o seu sistema operacional. Os seguintes comandos são para instalar o Cloud SDK no Ubuntu. Veja <https://cloud.google.com/sdk/docs/install-sdk#deb> para quaisquer atualizações neste procedimento.

O primeiro passo é adicionar o URI da CLI gcloud como uma fonte para pacotes:

```
echo "deb [signed-by=/usr/share/keyrings/cloud.google.gpg] https://packages.cloud.google.com/apt cloud-sdk main" | sudo tee -a /etc/apt/sources.list.d/google-cloud-sdk.list
```

Você também precisa importar a chave pública do Google Cloud, o que você faz com este comando:

```
curl https://packages.cloud.google.com/apt/doc/apt-key.gpg | sudo apt-key --keyring /usr/share/keyrings/cloud.google.gpg add -
```

Finalmente, você precisa atualizar a lista de pacotes do apt-get e depois usar apt-get para instalar o Cloud SDK:

```
sudo apt-get update && sudo apt-get install google-cloud-cli
```

Agora que o Cloud SDK está instalado, você pode executar comandos usando-o. O primeiro passo é inicializar o Cloud SDK usando o comando gcloud init, conforme mostrado aqui:

```
gcloud init
```

Quando você receber um link de autenticação, copie-o para o seu navegador. Será solicitado que você se autentique com o Google ao acessar essa URL. Em seguida, um código de resposta aparecerá em seu navegador. Copie isso para a janela do seu terminal e cole-o em resposta ao prompt que deve aparecer.

Em seguida, será solicitado que você insira um projeto. Se projetos já existirem em sua conta, eles serão listados. Você também tem a opção de criar um novo projeto neste ponto. O projeto que você selecionar ou criar será o projeto padrão usado ao emitir comandos por meio do Cloud SDK.

Criando uma Máquina Virtual com o Cloud SDK

Para criar uma VM a partir da linha de comando, você usará o comando gcloud. Você usa esse comando para muitas tarefas de gerenciamento na nuvem, incluindo trabalhar com os seguintes serviços:

- Compute Engine
- Instâncias do Cloud SQL
- Kubernetes Engine
- Cloud Dataproc
- Cloud DNS
- Cloud Deployment Manager

O comando gcloud é organizado em uma hierarquia de grupos, como o grupo compute para comandos do Compute Engine. Discutiremos outros grupos em capítulos posteriores; o foco aqui está no Compute Engine.

Um comando gcloud típico começa com o grupo, conforme mostrado aqui:

```
gcloud compute
```

Um subgrupo é usado nos comandos do Compute Engine para indicar com que tipo de recurso de computação você está trabalhando. Para criar uma instância, você usa este comando:

```
gcloud compute instances
```

E a ação que você deseja tomar é criar uma instância, então você usa isso:

```
gcloud compute instances create ace-instance-1 ace-instance-2
```

Se você não especificar parâmetros adicionais, como uma zona, o Google Cloud usará suas informações do seu projeto padrão. Você pode visualizar as informações do seu projeto usando o seguinte comando gcloud:

```
gcloud compute project-info describe
```

Para criar uma VM na zona us-central1-a, adicione o parâmetro de zona assim:

```
gcloud compute instances create ace-instance-1 ace-instance-2 --zone=us-central1-a
```

Você pode listar as VMs que você criou usando isso:

```
gcloud compute instances list
```

Os seguintes são parâmetros comumente usados com o comando de criação de instância:

■■ --boot-disk-size é o tamanho do disco de inicialização para um novo disco. O tamanho do disco pode estar entre 10 GB e 2 TB.

■■ --boot-disk-type é o tipo de disco. Use gcloud compute disk-types list para uma lista de tipos de disco disponíveis na zona em que a VM está sendo criada.

■■ --labels é a lista de pares chave-valor no formato de CHAVE=VALOR.

■■ --machine-type é o tipo de máquina a ser usado. Se não especificado, usa n1-standard-1. Use gcloud compute machine-types list para ver uma lista de tipos de máquinas disponíveis na zona que você está usando.

■■ --preemptible, se incluído, especifica que a VM será preemptível.

Para parâmetros adicionais, veja a documentação do gcloud compute instance create em <https://cloud.google.com/sdk/gcloud/reference/compute/instances/create>.

Para criar uma VM padrão com 8 CPUs e 30 GB de memória, você pode especificar e2-standard-2 como o tipo de máquina:

```
gcloud compute instances create ace-instance-n1s8 --machine-type=e2-standard-2
```

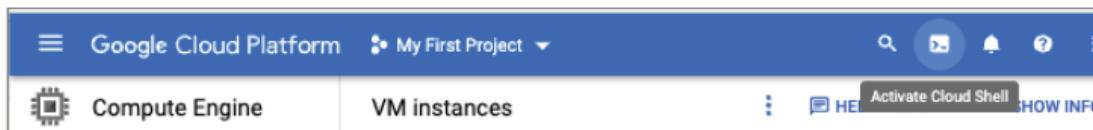
Se você quiser tornar esta instância preemptível, adicione o parâmetro preemptible:

```
gcloud compute instances create --machine-type=n1-standard-8 --preemptible ace-instance-1
```

Criando uma Máquina Virtual com o Cloud Shell

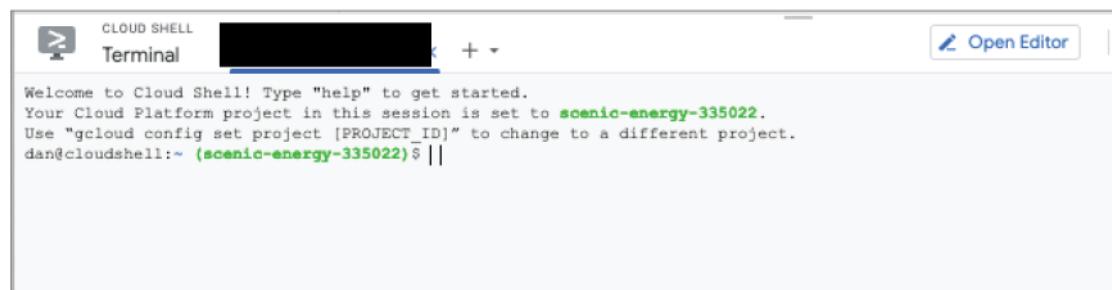
Uma alternativa para executar comandos gcloud localmente é executá-los em uma instância na nuvem. O Cloud Shell fornece essa capacidade. Para usar o Cloud Shell, inicie-o a partir do Cloud Console clicando no ícone do shell no canto superior direito do navegador, conforme mostrado na Figura 5.18.

FIGURE 5.18 Cloud Shell is activated through Cloud Console.



O Cloud SDK está instalado e o Cloud Shell fornece uma linha de comando Linux, conforme mostrado na Figura 5.19. Todos os comandos gcloud que você pode inserir no seu dispositivo local com o Cloud SDK instalado podem ser usados no Cloud Shell.

FIGURE 5.19 Cloud Shell opens a command-line window in the browser.



Gerenciamento Básico de Máquinas Virtuais

Quando as VMs estão em execução, você pode realizar tarefas básicas de gerenciamento usando o console ou comandos gcloud.

Iniciando e Parando Instâncias

No console, você pode visualizar uma lista de instâncias selecionando Compute Engine e depois VM Instances no painel lateral esquerdo do console. Você pode então selecionar uma VM para operar e listar opções de comando clicando nos ícones de reticências à direita. A Figura 5.20 mostra um exemplo.

Observe que você pode iniciar uma instância parada usando o comando de início que é habilitado no pop-up para instâncias paradas.

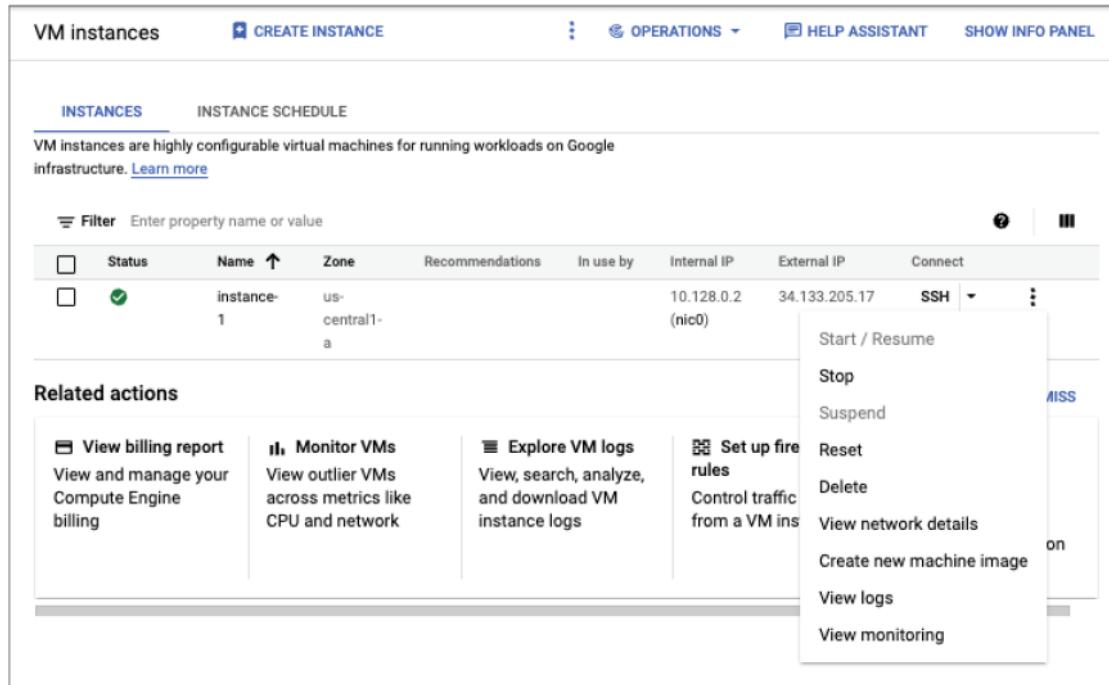
Você também pode usar gcloud para parar uma instância com o seguinte comando, onde INSTANCE-NAME é o nome da instância:

```
gcloud compute instances stop INSTANCE-NAME
```

Acesso à Rede para Máquinas Virtuais

Como engenheiro de nuvem, às vezes você precisará fazer login em uma VM para realizar algumas tarefas de administração. A maneira mais comum é usar SSH ao fazer login em um servidor Linux ou usar o Protocolo de Área de Trabalho Remota (RDP) ao fazer login em um servidor Windows.

FIGURE 5.20 Basic operations on VMs can be performed using a pop-up menu in the console.



A Figura 5.21 mostra o conjunto de opções para usar SSH a partir do console. Esta lista de opções aparece quando você clica no botão SSH associado a uma VM.

FIGURE 5.21 From the console, you can start an SSH session to log into a Linux server.

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
Green checkmark	instance-1	us-central1-a			10.128.0.2	34.133.205.17	SSH

Related actions

- View billing report
- Monitor VMs
- View outlier VMs across metrics like CPU and network
- View gcloud command
- View and instance logs
- Use another SSH client
- from a VM instance
- updates and view patch compliance

Escolher a opção "Abrir em uma janela do navegador" abrirá uma nova janela do navegador e exibirá uma janela de terminal para acessar a linha de comando no servidor, conforme mostrado na Figura 5.22.

FIGURE 5.22 A terminal window opens in a new browser window when using SSH-in-browser.



Monitoramento de uma Máquina Virtual

Enquanto sua VM estiver em execução, você pode monitorar a CPU, disco e carga de rede visualizando a aba de Monitoramento na página de Detalhes da Instância da VM.

Para acessar informações de monitoramento no console, selecione uma instância de VM na página de Instância da VM clicando no nome da VM que você deseja monitorar. Isso exibirá a página de Detalhes da VM. Selecione a opção de Monitoramento perto do topo da página para visualizar detalhes de monitoramento.

A Figura 5.23 mostra as informações exibidas sobre a utilização da CPU, rede e operações de disco.

Custo de Máquinas Virtuais

Parte do gerenciamento básico de uma VM é rastrear os custos das instâncias que você está executando. Se você deseja rastrear custos automaticamente, pode habilitar a cobrança do Cloud e configurar a Exportação de Cobrança. Isso produzirá relatórios diários sobre o uso e custo das VMs.

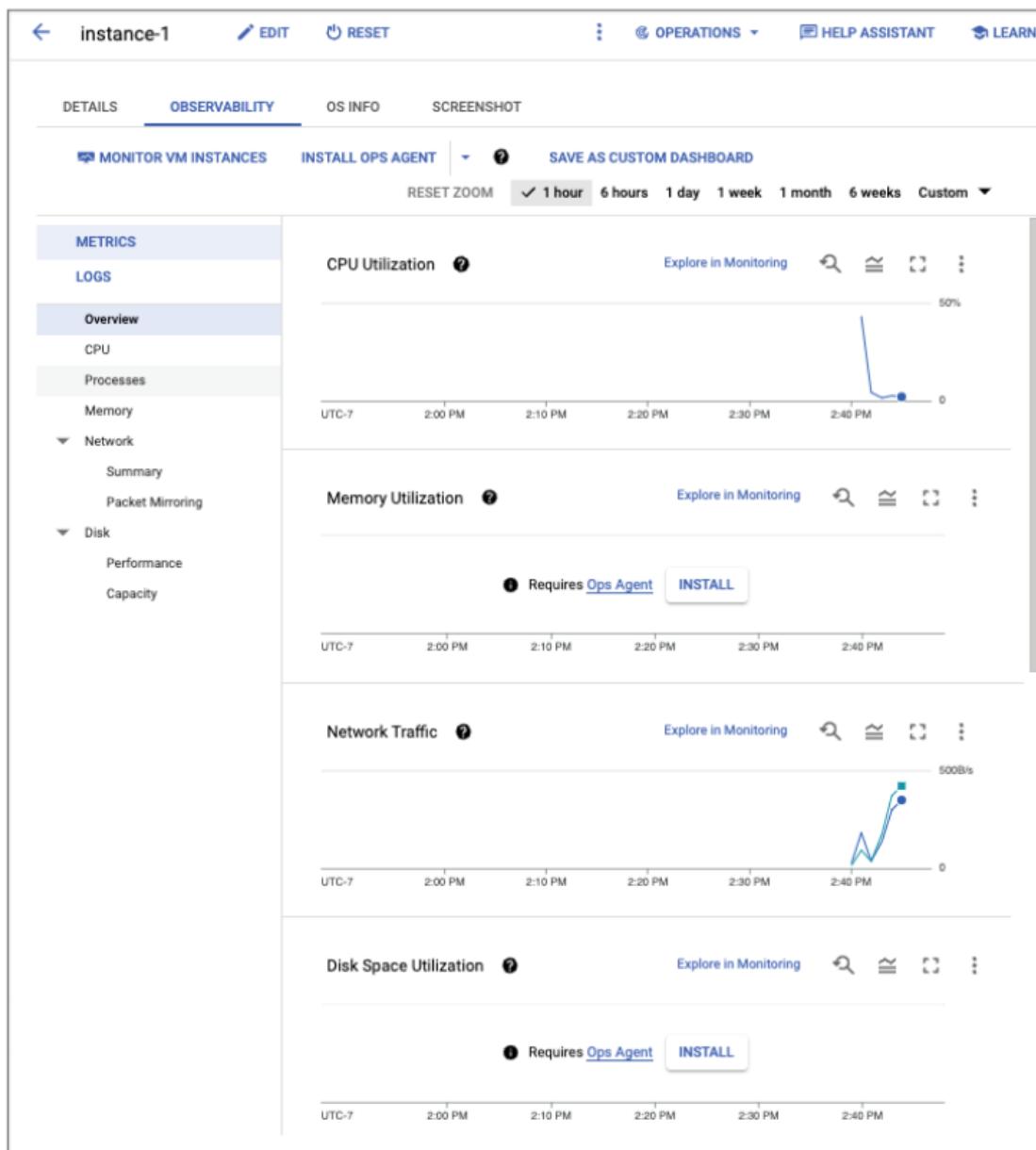
Aqui estão as coisas mais importantes a lembrar sobre os custos das VMs:

- VMs são cobradas em incrementos de 1 segundo.
- O custo é baseado no tipo de máquina. Quanto mais CPUs e memória utilizadas, maior o custo.
- O Google oferece descontos para uso sustentado, mas descontos não são oferecidos em todos os tipos de instância.
- VMs são cobradas por um mínimo de 1 minuto de uso.
- VMs spot podem economizar até 80% do custo de uma VM.

Para informações adicionais sobre preços, veja <https://cloud.google.com/compute/vm-instance-pricing>.

FIGURA 5.23 A aba de Observabilidade da página de Detalhes da Instância da VM

FIGURE 5.23 The Observability tab of the VM Instance Details page



Diretrizes para Planejamento, Implantação e Gerenciamento de Máquinas Virtuais

Considere as seguintes diretrizes para ajudar a otimizar seu trabalho com VMs. Essas diretrizes aplicam-se ao trabalho com um pequeno número de VMs. Capítulos posteriores fornecerão diretrizes adicionais para trabalhar com clusters e grupos de instâncias, que são conjuntos de VMs configuradas de forma similar.

- Escolha um tipo de máquina com o menor número de CPUs e a menor quantidade de memória que ainda atenda às suas necessidades, incluindo capacidade máxima. Isso minimizará o custo da VM.

- Use o console para administração ad hoc de VMs. Use scripts com comandos gcloud para tarefas que serão repetidas.

- Use scripts de inicialização para realizar atualizações de software e outras tarefas que devem ser realizadas na inicialização.
- Se você fizer várias modificações em uma imagem de máquina, considere salvá-la e usá-la com novas instâncias em vez de executar o mesmo conjunto de modificações em cada nova instância.
- Se você pode tolerar interrupções não planejadas, use VMs spot para reduzir custos.
- Use SSH ou RDP para acessar uma VM para realizar tarefas no nível do sistema operacional.
- Use o Cloud Console, Cloud Shell ou Cloud SDK para realizar tarefas no nível da VM.

Resumo

O Google Cloud Console é uma interface gráfica de usuário baseada na web para gerenciar recursos do Google Cloud. O Cloud SDK é um pacote de linha de comando que permite aos engenheiros gerenciar recursos da nuvem a partir da linha de comando de seu dispositivo local. O Cloud Shell é uma interface de terminal baseada na web para VMs. O Cloud SDK está instalado no Cloud Shell.

Ao criar uma VM, você deve especificar uma série de parâmetros, incluindo um nome para a VM, uma região e zona onde a VM será executada, um tipo de máquina que especifica o número de vCPUs e a quantidade de memória, e um disco de inicialização que inclui um sistema operacional.

`gcloud` é o comando de nível superior da estrutura de comando hierárquico no Cloud SDK. Tarefas comuns ao gerenciar VMs incluem iniciar e parar instâncias, usar SSH para acessar um terminal na VM, monitoramento e rastreamento do custo da VM.

Essenciais para o Exame

Entenda como usar o Cloud Console e o Cloud SDK para criar, iniciar e parar VMs. Parâmetros que você precisará fornecer ao criar uma VM incluem nome, tipo de máquina, região, zona e disco de inicialização. Entenda a necessidade de criar uma VM em um projeto.

Saiba como configurar uma VM spot usando o Cloud Console e os comandos `gcloud`. Saiba quando usar uma VM spot e quando não usar. Saiba que VMs spot custam até 80% menos que VMs padrão.

Conheça o propósito das opções avançadas, incluindo VMs Blindadas (Shielded VMs) e configurações avançadas de disco de inicialização. Saiba que opções avançadas fornecem segurança adicional. Entenda os tipos de proteções fornecidas.

Saiba como usar comandos `gcloud compute instance` para listar, iniciar e parar VMs. Conheça a estrutura dos comandos `gcloud`. Comandos `gcloud` começam com `gcloud` seguido por um serviço, como `compute`, seguido por um tipo de recurso, como `instances`, seguido por um comando ou verbo, como `create`, `list` ou `describe`.

Entenda como monitorar uma VM. Saiba onde encontrar a utilização da CPU, monitoramento de rede e monitoramento de disco nas páginas de Instâncias da VM no console. Conheça a diferença entre listar e descrever instâncias com um comando gcloud.

Conheça os fatores que determinam o custo de uma VM. Saiba que o Google cobra pelo segundo com um mínimo de 1 minuto. Entenda que os custos de um tipo de máquina podem ser diferentes em diferentes localizações. Saiba que o custo é baseado no número de vCPUs e memória.

Questões

1. Você acabou de abrir o console do Google Cloud em <http://console.google.com>. Você se autenticou com o usuário que deseja usar. Qual é uma das primeiras coisas que você deve fazer antes de realizar tarefas em VMs?
 - A. Abrir o Cloud Shell.
 - B. Verificar se você pode fazer login em uma VM usando SSH.
 - C. Verificar se o projeto selecionado é aquele com o qual você deseja trabalhar.
 - D. Revisar a lista de VMs em execução.
2. Qual é uma tarefa única que você precisará concluir antes de usar o console?
 - A. Configurar a cobrança.
 - B. Criar um projeto.
 - C. Criar um bucket de armazenamento.
 - D. Especificar uma zona padrão.
3. Um colega pediu sua assistência para configurar um ambiente de teste no Google Cloud. Eles nunca trabalharam no Google Cloud. Você sugere começar com uma única VM. Qual das seguintes é o conjunto mínimo de informações que você precisará?
 - A. Um nome para a VM e um tipo de máquina.
 - B. Um nome para a VM, um tipo de máquina, uma região e uma zona.
 - C. Um nome para a VM, um tipo de máquina, uma região, uma zona e um bloco CIDR.
 - D. Um nome para a VM, um tipo de máquina, uma região, uma zona e um endereço IP.
4. Um arquiteto sugeriu um tipo de máquina específico para sua carga de trabalho. Você está no console criando uma VM e não vê o tipo de máquina na lista de tipos de máquina disponíveis. Qual pode ser a razão para isso?
 - A. Você selecionou a sub-rede incorreta.
 - B. Esse tipo de máquina não está disponível na zona que você especificou.
 - C. Você escolheu um sistema operacional incompatível.
 - D. Você não especificou uma configuração de memória correta.
5. Seu gerente pede sua ajuda para entender os custos de computação em nuvem. Sua equipe executa dezenas de VMs para três aplicações diferentes. Duas das aplicações são para uso pelo departamento de marketing e uma é usada pelo departamento financeiro. Seu gerente quer uma maneira de cobrar cada

departamento pelo custo das VMs usadas para suas aplicações. O que você sugeriria para ajudar a resolver esse problema?

- A. Controles de acesso
 - B. Discos persistentes
 - C. Etiquetas e descrições
 - D. Apenas descrições
6. Se você quisesse definir a propriedade preemptível usando o Cloud Console, em qual seção da página Criar Uma Instância você encontraria a opção?
- A. Política de Disponibilidade
 - B. Identidade e Acesso à API
 - C. Tenência Única
 - D. Rede
7. Você precisa configurar um servidor com um alto nível de segurança. Você quer estar preparado em caso de ataques ao seu servidor por alguém tentando injetar um rootkit (um tipo de malware que pode alterar o sistema operacional). Qual opção você deve selecionar ao criar uma VM?
- A. Firewall
 - B. VM Blindada (Shielded VM)
 - C. Chaves SSH em toda a organização
 - D. Serviço de controle de integridade do disco de inicialização
8. Todos os seguintes parâmetros podem ser definidos ao adicionar um disco adicional pelo Google Cloud Console, exceto:
- A. Tipo de disco
 - B. Gerenciamento de chave de criptografia
 - C. Tamanho do bloco
 - D. Imagem fonte para o disco
9. Você lidera uma equipe de engenheiros de nuvem que mantêm recursos de nuvem para vários departamentos em sua empresa. Você notou um problema com a deriva de configuração. Algumas configurações de máquina não estão mais no mesmo estado que estavam quando criadas. Você não consegue encontrar anotações ou documentação sobre como as mudanças foram feitas ou por quê. Qual prática você implementaria para resolver esse problema?
- A. Fazer com que todos os engenheiros de nuvem usem apenas a interface de linha de comando no Cloud Shell.

- B. Escrever scripts usando comandos gcloud para mudar a configuração e armazenar esses scripts em um sistema de controle de versão.
- C. Fazer anotações ao fazer mudanças na configuração e armazená-las no Google Drive.
- D. Limitar privilégios para que apenas você possa fazer mudanças, e você sempre saberá quando e por que as configurações foram mudadas.
10. Quando você está usando a interface de linha de comando do Cloud SDK, qual dos seguintes faz parte dos comandos para administrar recursos no Compute Engine?
- A. gcloud compute instances
 - B. gcloud instances
 - C. gcloud instances compute
 - D. Nenhuma das opções acima
11. Um engenheiro de nuvem recém-contratado está tentando entender quais VMs (Máquinas Virtuais) estão rodando em um projeto específico. Como o engenheiro poderia obter informações resumidas sobre cada VM rodando no projeto?
- A. Execute o comando gcloud compute list.
 - B. Execute o comando gcloud compute instances list.
 - C. Execute o comando gcloud instances list.
 - D. Execute o comando gcloud list instances.
12. Ao criar uma VM usando a linha de comando, como você deve especificar rótulos para a VM?
- A. Use a opção --labels com rótulos no formato de CHAVES:VALORES.
 - B. Use a opção --labels com rótulos no formato de CHAVES=VALOR.
 - C. Use a opção --labels com rótulos no formato de CHAVES,VALORES.
 - D. Isso não é possível na linha de comando.
13. Na configuração avançada do disco de inicialização, quais operações você pode especificar ao criar uma nova VM?
- A. Adicionar um novo disco, reformatar um disco existente, anexar um disco existente.
 - B. Adicionar um novo disco e reformatar um disco existente.
 - C. Adicionar um novo disco e anexar um disco existente.
 - D. Reformatar um disco existente e anexar um disco existente.

14. Você adquiriu um conjunto de dados de 10 GB de uma empresa de pesquisa terceirizada. Um grupo de cientistas de dados gostaria de acessar esses dados a partir de seus programas de estatísticas escritos em R. R funciona bem com sistemas de arquivos Linux e Windows, e os cientistas de dados estão familiarizados com operações de arquivo em R. Os cientistas de dados gostariam de ter sua própria VM dedicada com os dados disponíveis no sistema de arquivos da VM. Qual é uma maneira de tornar esses dados prontamente disponíveis em uma VM e minimizar as etapas que os cientistas de dados terão que realizar?

- A. Armazene os dados no Cloud Storage.
- B. Crie VMs usando uma imagem-fonte criada a partir de um disco com os dados nele.
- C. Armazene os dados no Google Drive.
- D. Carregue os dados no BigQuery.

15. A aba de Rede do formulário de Criação de VM é onde você realizaria qual das seguintes operações?

- A. Definir o endereço IP da VM.
- B. Adicionar uma interface de rede à VM.
- C. Especificar um roteador padrão.
- D. Alterar regras de configuração de firewall.

16. Você quer criar uma VM usando o comando gcloud. Que parâmetro você incluiria para especificar o tipo de disco de inicialização?

- A. boot-disk-type
- B. boot-disk
- C. disk-type
- D. type-boot-disk

17. Qual dos seguintes comandos criará uma VM com quatro CPUs que se chama web-server-1?

- A. gcloud compute instances create --machine-type=n1-standard-4 web-server-1
- B. gcloud compute instances create --cpus=4 web-server-1
- C. gcloud compute instances create --machine-type=n1-standard-4 --instance-name=web-server-1
- D. gcloud compute instances create --machine-type=n1-4-cpu web-server-1

18. Qual dos seguintes comandos irá parar uma VM chamada web-server-1?

- A. gcloud compute instances halt web-server-1
- B. gcloud compute instances --terminate web-server1

- C. gcloud compute instances stop web-server-1
 - D. gcloud compute stop web-server-1
19. Você acabou de criar uma VM Ubuntu e quer fazer login na VM para instalar alguns pacotes de software. Qual serviço de rede você usaria para acessar a VM?
- A. FTP
 - B. SSH
 - C. RDP
 - D. ipconfig
20. Sua equipe de gestão está considerando três diferentes provedores de nuvem. Você foi solicitado a resumir informações de faturamento e custo para ajudar a equipe de gestão a comparar as estruturas de custo entre as nuvens. O que você mencionaria sobre o custo de VMs no Google Cloud?
- A. VMs são faturadas em incrementos de 1 segundo, o custo varia com o número de CPUs e quantidade de memória em um tipo de máquina, você pode criar tipos de máquina personalizados, VMs preemptivas custam até 80 por cento menos que VMs padrão, e o Google oferece descontos para uso sustentado.
 - B. VMs são faturadas em incrementos de 1 segundo e VMs podem rodar até 24 horas antes de serem desligadas.
 - C. O Google oferece descontos para uso sustentado apenas em algumas regiões, o custo varia com o número de CPUs e quantidade de memória em um tipo de máquina, você pode criar tipos de máquina personalizados, VMs preemptivas custam até 80 por cento menos que VMs padrão.
 - D. VMs são cobradas por um mínimo de 1 hora de uso, e o custo varia com o número de CPUs e quantidade de memória em um tipo de máquina.

CAPÍTULO 6

Gerenciamento de Máquinas Virtuais

ESTE CAPÍTULO COBRE OS SEGUINTESS OBJETIVOS DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 4.1 Gerenciamento de recursos do Compute Engine

Após criar máquinas virtuais, você precisará trabalhar tanto com instâncias únicas de máquinas virtuais (VMs) quanto com grupos de VMs que executam a mesma configuração. Estes últimos são chamados de grupos de instâncias e são introduzidos neste capítulo.

Este capítulo começa com uma descrição das tarefas comuns de gerenciamento e como completá-las no console, seguido por uma descrição de como completá-las no Cloud Shell ou com a linha de comando do Cloud SDK. Em seguida, você aprenderá a configurar e gerenciar grupos de instâncias. O capítulo conclui com uma discussão sobre diretrizes para gerenciar VMs.

Gerenciamento de Instâncias Únicas de Máquina Virtual

Começamos discutindo como gerenciar uma única instância de uma VM. Por instância única, queremos dizer uma criada por si só e não em um grupo de instâncias ou outro tipo de cluster. Recorde dos capítulos anteriores que existem três maneiras de trabalhar com instâncias: no Cloud Console, no Cloud Shell e com a linha de comando do Cloud SDK. Tanto o Cloud Shell quanto a linha de comando do Cloud SDK fazem uso de comandos `gcloud`, então descreveremos o Cloud Shell e o Cloud SDK juntos nesta seção.

Gerenciando Instâncias Únicas de Máquina Virtual no Console

As tarefas básicas de gerenciamento de VM com as quais você deve estar familiarizado são criar, parar e deletar instâncias. Cobrimos a criação de instâncias no capítulo anterior, então focaremos nas outras tarefas aqui. Você também deve estar familiarizado com listar VMs, anexar unidades de processamento gráfico (GPUs) às VMs e trabalhar com snapshots e imagens.

Iniciando, Parando e Deletando Instâncias

Para começar a trabalhar, abra o console e selecione Compute Engine. Em seguida, selecione Instâncias de VM. Isso exibirá uma janela como a da Figura 6.1, mas com VMs diferentes listadas. Neste exemplo, existem três VMs.

FIGURE 6.1 The VM Instance panel in the Compute Engine section of Cloud Console

The screenshot shows the 'VM instances' panel in the Compute Engine section of the Cloud Console. At the top, there are buttons for 'CREATE INSTANCE', 'IMPORT VM', 'OPERATIONS', 'HELP ASSISTANT', 'SHOW INFO PANEL', and 'LEARN'. Below this, there are tabs for 'INSTANCES' and 'INSTANCE SCHEDULE'. A note states: 'VM instances are highly configurable virtual machines for running workloads on Google infrastructure. [Learn more](#)'. A 'Filter' input field is present. The main table lists three instances:

<input type="checkbox"/>	Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	✓	instance-1	us-central1-a			10.128.0.2 (nic0)	34.133.205.17	SSH <input type="button" value="⋮"/>
<input type="checkbox"/>	✓	instance-2	us-central1-a			10.128.0.3 (nic0)	35.202.99.211	SSH <input type="button" value="⋮"/>
<input type="checkbox"/>	✓	instance-3	us-central1-a			10.128.0.4 (nic0)	104.154.134.53	SSH <input type="button" value="⋮"/>

Source: Google LLC

As três instâncias na Figura 6.1 estão todas em execução. Você pode parar as instâncias clicando no ícone de três pontos no lado direito da linha que lista os atributos da VM. Essa ação exibe uma lista de comandos. A Figura 6.2 mostra a lista de comandos disponíveis para a instância-1.

FIGURE 6.2 The list of commands available from the console for changing the state of a VM

The screenshot shows a list of Compute Engine VM instances. There are three instances listed: 'instance-1' (Status: Running, Zone: us-central1-a), 'instance-2' (Status: Running, Zone: us-central1-a), and 'instance-3' (Status: Running, Zone: us-central1-a). Each instance has columns for Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, and Connect (SSH dropdown). To the right of the list is a vertical menu with options: Start / Resume, Stop, Suspend, Reset, Delete, View network details, Create new machine image, View logs, and View monitoring. A 'DISMISS' button is at the bottom of this menu.

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	instance-1	us-central1-a			10.128.0.2 (nic0)	34.133.205.17	SSH
<input type="checkbox"/>	instance-2	us-central1-a			10.128.0.3 (nic0)	35.202.99.211	SSH
<input type="checkbox"/>	instance-3	us-central1-a			10.128.0.4 (nic0)	104.154.134.53	SSH

Related actions

- View billing report
- View and manage your Compute Engine billing
- Monitor VMs
- View outlier VMs across metrics like CPU and network
- Explore VM logs
- View, search, analyze, and download VM instance logs
- Set up firewall
- Control traffic to and from a VM instance

Actions

- Start / Resume
- Stop
- Suspend
- Reset
- Delete
- View network details
- Create new machine image
- View logs
- View monitoring

DISMISS

Source: Google LLC

Se você selecionar Parar no menu de comandos, a instância será interrompida. Quando uma instância é interrompida, ela não está consumindo recursos de computação, então você não será cobrado. A instância ainda existe e pode ser iniciada novamente quando você precisar. A Figura 6.3 mostra um formulário de aviso indicando que você está prestes a parar uma VM. Você pode clicar na caixa de diálogo no canto inferior esquerdo para suprimir essa mensagem.

Quando você para uma VM, a marca de verificação verde à esquerda muda para um círculo cinza com um quadrado branco, e a opção SSH é desativada, conforme mostrado na Figura 6.4.

FIGURE 6.3 A warning message that may appear about stopping a VM

The screenshot shows the same VM list as Figure 6.2, but with a modal dialog box centered over the 'instance-1' row. The dialog title is 'Stop instance-1?' and contains the text: 'Stop shuts down the instance. If the shutdown doesn't complete within 90 seconds, the instance is forced to halt. This can lead to file-system corruption. Do you want to stop instance "instance-1"?'. At the bottom of the dialog are 'CANCEL' and 'STOP' buttons. The 'STOP' button is highlighted in blue. The background list shows 'instance-1' with a greyed-out status and a disabled SSH dropdown, while 'instance-2' and 'instance-3' remain active.

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input checked="" type="checkbox"/>	instance-1	us-central1-a			10.128.0.2 (nic0)	34.133.205.17	SSH
<input type="checkbox"/>	instance-2	us-central1-a			10.128.0.3 (nic0)	35.202.99.211	SSH
<input type="checkbox"/>	instance-3	us-central1-a			10.128.0.4 (nic0)	104.154.134.53	SSH

Related actions

- View billing report
- View and manage your Compute Engine billing
- Stop instance-1?
- Set up firewall rules
- Control traffic to and from a VM instance
- Patch management
- Schedule patch updates and view patch compliance on VM instances

Actions

- Start / Resume
- Stop**
- Suspend
- Reset
- Delete
- View network details
- Create new machine image
- View logs
- View monitoring

DISMISS

Source: Google LLC

FIGURE 6.4 When VMs are stopped, the icon on the left changes and SSH is no longer available.

VM Instances							
	Status	Name	Zone	Recommendations	In use by	Internal IP	External IP
<input type="checkbox"/>	instance-1	us-central1-a				10.128.0.2 (nic0)	None
<input checked="" type="checkbox"/>	instance-2	us-central1-a				10.128.0.3 (nic0)	35.202.99.211
<input checked="" type="checkbox"/>	instance-3	us-central1-a				10.128.0.4 (nic0)	104.154.134.53

Source: Google LLC

Para iniciar uma VM parada, clique no ícone de três pontos à direita para exibir o menu de comandos disponíveis. Observe na Figura 6.5 que Iniciar agora está disponível, mas Parar e Reiniciar não estão. O comando Reiniciar reinicia uma VM. As propriedades da VM não mudarão, mas os dados na memória serão perdidos.

Quando uma VM é reiniciada, o conteúdo da memória é perdido. Se você precisar preservar dados entre reinicializações ou para uso em outras VMs, salve os dados em um disco persistente ou no Cloud Storage.

Quando você terminou com uma instância e não precisa mais dela, você pode excluí-la. Excluir uma VM a remove do Cloud Console e libera recursos, como o armazenamento usado para manter a imagem da VM quando parada. Excluir uma instância do Cloud Console exibirá uma mensagem de aviso, mostrada na Figura 6.6.

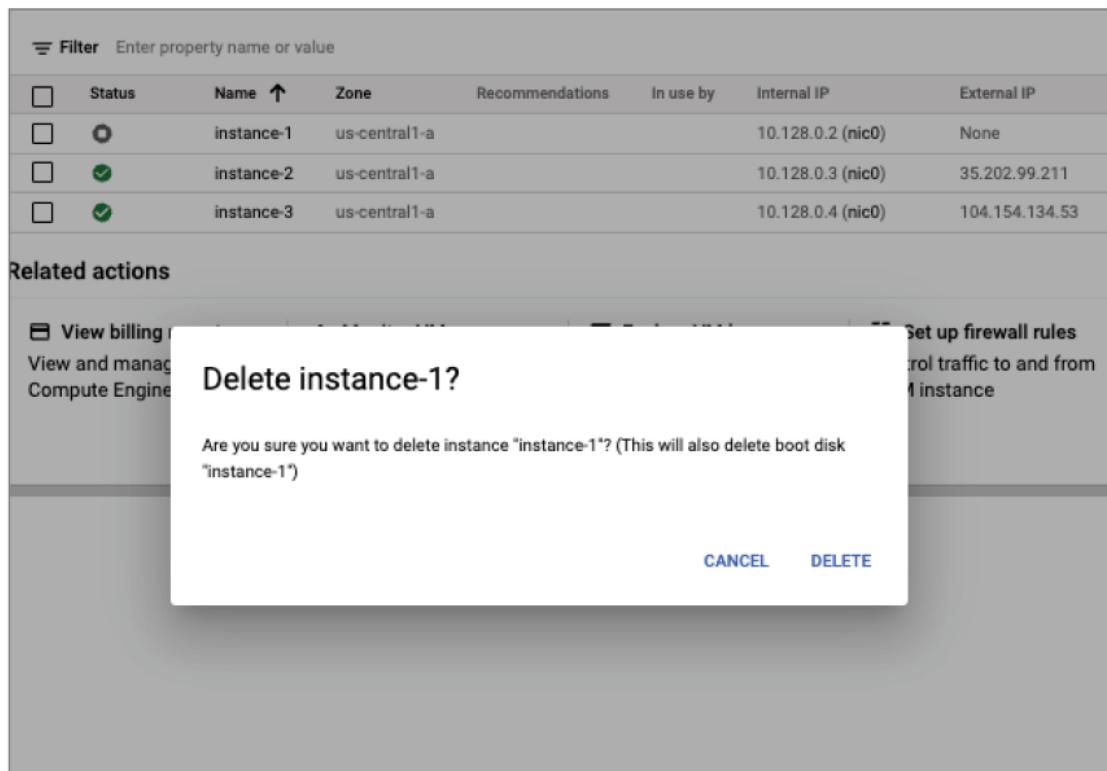
FIGURE 6.5 When VMs are stopped, Stop and Reset are no longer available, but Start / Resume is available as a command.

VM Instances							
	Status	Name	Zone	Recommendations	In use by	Internal IP	External IP
<input type="checkbox"/>	instance-1	us-central1-a				10.128.0.2 (nic0)	None
<input checked="" type="checkbox"/>	instance-2	us-central1-a				10.128.0.3 (nic0)	35.202.99.211
<input checked="" type="checkbox"/>	instance-3	us-central1-a				10.128.0.4 (nic0)	104.154.134.53

Related actions

- Start / Resume**
- Stop
- Suspend
- Reset
- Delete
- View network details
- Create new machine image
- View logs
- View monitoring

FIGURE 6.6 Deleting an instance from the console will display a warning message such as this.



Source: Google LLC

Visualização do Inventário de Máquinas Virtuais

A página de Instâncias de VM do Cloud Console mostrará uma lista de VMs, se existirem no projeto atual. Se você tem um grande número de instâncias, pode ser útil filtrar a lista para ver apenas instâncias de interesse. Faça isso usando a caixa Filtrar Instâncias de VM acima da lista de VMs, conforme mostrado na Figura 6.7.

FIGURE 6.7 List of instances filtered by search criteria

Filter <input type="text" value="instance-2"/> Enter property name or value						
Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP
<input type="checkbox"/>	instance-2	us-central1-a				

Source: Google LLC

Neste exemplo, especificamos que queremos ver apenas a instância chamada instance-2.

Além de especificar nomes de instâncias, você também pode filtrar pelo seguinte:

■■ Etiquetas

■■ IP interno

- IP externo
- Status
- Zona
- Rede
- Proteção contra exclusão

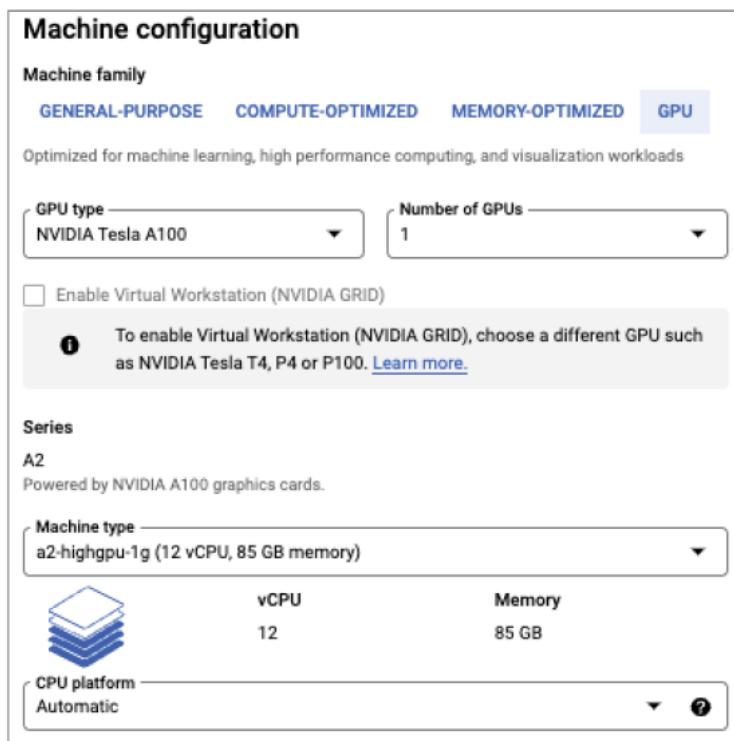
Se você definir várias condições de filtro, todas devem ser verdadeiras para que uma VM seja listada, a menos que você declare explicitamente o operador OU.

Anexando GPUs a uma Instância

GPs são usadas para aplicações intensivas em matemática, como visualizações e aprendizado de máquina. GPUs realizam cálculos matemáticos e permitem que algum trabalho seja transferido da CPU para a GPU. O Compute Engine tem uma família de máquinas especificamente projetada para VMs com GPUs. Para usar GPUs, você também precisará instalar drivers de GPU ou usar uma imagem que já tenha drivers de GPU instalados.

Ao criar uma instância no console, você pode escolher a família de máquinas GPU para ver as opções para trabalhar com GPUs. (Veja a Figura 6.8.)

FIGURE 6.8 GPU machine family supports a variety of GPU types, and a number of GPUs and CPU platforms.



Source: Google LLC

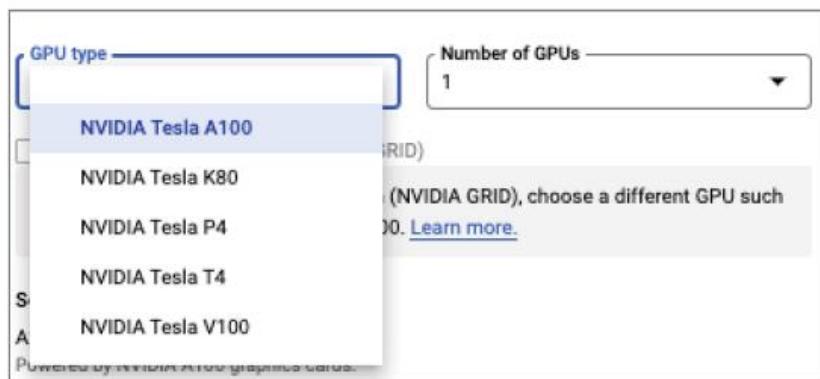
Para adicionar uma GPU a uma instância, você deve iniciar uma instância na qual as bibliotecas de GPU tenham sido instaladas ou serão instaladas. Por exemplo, você pode

usar uma das imagens do Google Cloud que tenha bibliotecas de GPU instaladas, incluindo as imagens de Aprendizado Profundo, conforme mostrado na Figura 6.8.

Você também deve verificar se a instância será executada em uma zona que tenha GPUs disponíveis.

Os parâmetros que você pode configurar incluem Tipo de GPU e Número de GPUs. A Figura 6.9 mostra algumas opções de GPU. O tipo de GPU determinará o número de GPUs disponíveis. Por exemplo, atualmente o NVIDIA Tesla A100 pode ser usado em configurações de 1, 2, 4, 8 ou 16 GPUs, enquanto o NVIDIA Tesla T4 pode ser usado em configurações de 1, 2 ou 4 GPUs.

FIGURE 6.9 Some GPU options available in Compute Engine



Source: Google LLC

Assim como em outras famílias de máquinas, você pode especificar um tipo de máquina. Você também pode especificar uma plataforma de CPU, como Intel Skylake ou posterior, ou Intel Ivy Bridge ou posterior. Essas são opções de microarquitetura. O Compute Engine automaticamente escolherá uma plataforma de CPU por padrão. Existem algumas restrições no uso de GPUs; por exemplo, GPUs não podem ser anexadas a máquinas de memória compartilhada. Para a documentação mais recente sobre restrições de GPUs e uma lista de zonas com GPUs, veja <https://cloud.google.com/compute/docs/gpus>.

Trabalhando com Snapshots

Snapshots são cópias de dados em um disco persistente. Você usa snapshots para salvar dados em um disco para que possa restaurá-los. Esta é uma maneira conveniente de fazer vários discos persistentes com os mesmos dados ou para fazer backup de um disco para que você possa recuperar o estado do disco em um determinado ponto no tempo.

Quando você cria um snapshot pela primeira vez, o Google Cloud fará uma cópia completa dos dados no disco persistente. Na próxima vez que você criar um snapshot desse disco, o Google Cloud copiará apenas os dados que mudaram desde o último snapshot. Isso otimiza o armazenamento enquanto mantém o snapshot atualizado com os dados que estavam no disco na última vez que uma operação de snapshot ocorreu.

Se você estiver executando um banco de dados ou outra aplicação que possa bufferizar dados na memória antes de escrever no disco, certifique-se de esvaziar os buffers do disco antes de criar o snapshot; caso contrário, dados na memória que deveriam ser escritos no disco podem ser perdidos. A maneira de esvaziar os buffers do disco variará de acordo com a aplicação. Por exemplo, o MySQL tem um comando FLUSH.

Para trabalhar com snapshots, um usuário deve ser atribuído ao papel de Administração de Armazenamento do Compute. Vá para a página de Gerenciamento de Identidade e Acesso (IAM), selecione Papéis e depois especifique o endereço de e-mail de um usuário para ser atribuído ao papel.

Para criar um snapshot do Cloud Console, exiba as opções do Compute Engine e selecione Snapshots no painel esquerdo, conforme mostrado na Figura 6.10.

Depois, clique em Criar Snapshot para exibir o formulário na Figura 6.11. Especifique um nome e, opcionalmente, uma descrição. Você também pode adicionar etiquetas ao snapshot. É uma boa prática etiquetar todos os recursos com uma convenção de rotulagem consistente. No caso de snapshots, as etiquetas podem indicar o tipo de dados no disco e a aplicação que usa os dados.

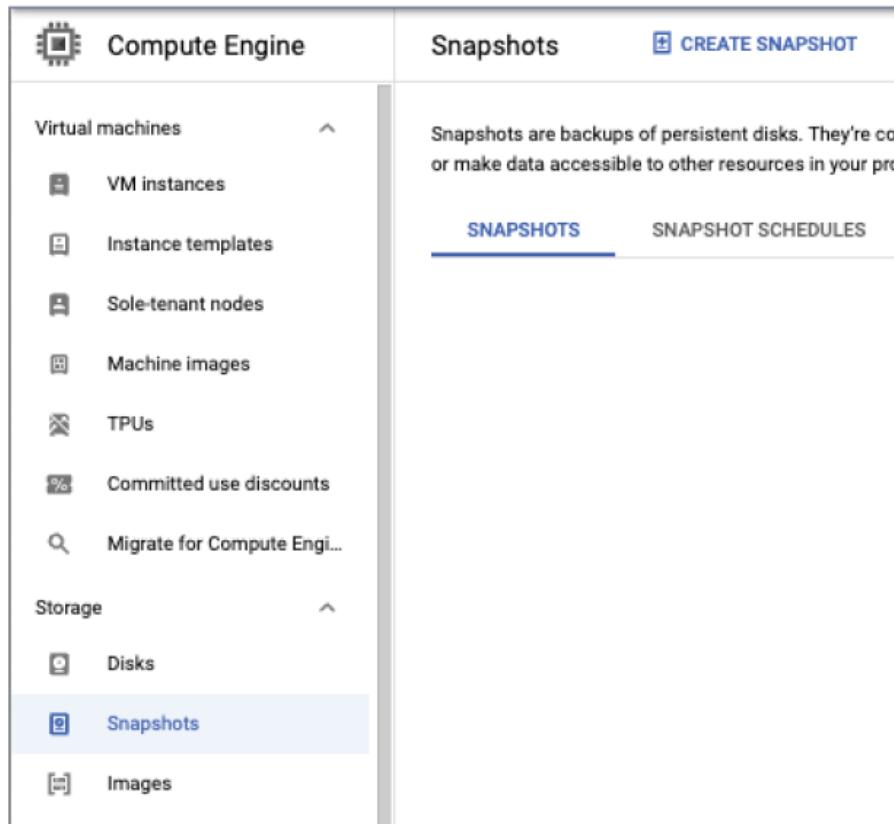
Você tem a opção de armazenar o snapshot regionalmente ou multirregionalmente.

Trabalhando com Imagens

Imagens são semelhantes a snapshots no sentido de que são cópias do conteúdo do disco. A diferença é que snapshots são usados para tornar os dados disponíveis em um disco, enquanto imagens são usadas para criar VMs. VMs também podem ser criadas a partir de snapshots, desde que esse snapshot seja feito a partir de um disco de inicialização. A principal diferença de armazenamento entre imagens e snapshots é que snapshots oferecem backups incrementais, enquanto imagens são um único backup completo. Imagens podem ser criadas a partir do seguinte:

- Disco
- Snapshot
- Imagem
- Arquivo de armazenamento em nuvem
- Disco virtual

FIGURE 6.10 Creating a snapshot using Cloud Console



Source: Google LLC

Para criar uma imagem, escolha a opção Imagem na página do Compute Engine no Cloud Console, conforme mostrado na Figura 6.12. Isso lista as imagens disponíveis.

Selecione Criar Imagem para mostrar o formulário na Figura 6.13. Este formulário permite que você crie uma nova imagem especificando um nome, descrição e etiquetas. Imagens têm um atributo opcional chamado Família, que permite agrupar imagens. Quando uma família é especificada, a imagem mais recente e não depreciada na família é usada.

O formulário fornece uma lista suspensa de opções para a fonte da imagem, conforme mostrado na Figura 6.14.

FIGURE 6.11 Form for creating a snapshot

The screenshot shows the 'Create a snapshot' interface. At the top, there's a back arrow and the title 'Create a snapshot'. Below the title, a descriptive text states: 'Snapshots are backups of persistent disks. They're commonly used to recover, transfer, or make data accessible to other resources in your project.' with a 'Learn more' link. The main form fields include:

- Name ***: A text input field containing 'snapshot-1'. Below it, a note says 'Name is permanent'.
- Description**: An empty text area for adding a description.
- Source disk ***: A dropdown menu currently set to 'Source disk'.
- Location ?**: A section with a note about network transfer fees and a 'Learn more' link. It includes two radio button options: 'Multi-regional' and 'Regional', with 'Regional' selected. A 'Select location' dropdown menu is also present.
- Labels ?**: A section with a '+ ADD LABEL' button.
- Note at the bottom**: 'Your free trial credit will be used for this snapshot. [GCP Free Tier](#)'.

Source: Google LLC

Quando você escolhe Imagem como o tipo de fonte, você pode escolher uma imagem do projeto atual ou de outros projetos (veja Figura 6.15).

Se você escolher um arquivo do Cloud Storage como fonte, você pode navegar pelo seu bucket do Cloud Storage para encontrar um arquivo para usar como fonte (veja Figura 6.16).

Depois de ter criado uma imagem, você pode excluí-la ou depreciá-la marcando a caixa ao lado do nome da imagem e selecionando Excluir ou Depreciar na linha de comandos acima da lista. Você pode excluir e depreciar apenas imagens personalizadas, não imagens fornecidas pelo Google Cloud.

FIGURE 6.12 Images available. From here, you can create additional images.

IMAGES		IMAGE IMPORT HISTORY		IMAGE EXPORT HISTORY					
<input type="checkbox"/>	Status	Name	Location	Archive size	Disk size	Created by	Family	Architecture	Actions
<input type="checkbox"/>	<input checked="" type="checkbox"/>	c0-deeplearning-common-cpu-v20221026-debian-10	asia, eu, us	—	50 GB	Debian	common-cpu-debian-10	—	⋮
<input type="checkbox"/>	<input checked="" type="checkbox"/>	c0-deeplearning-common-cu113-v20221026-debian-10	asia, eu, us	—	50 GB	Debian	common-dl-gpu-debian-10	—	⋮
<input type="checkbox"/>	<input checked="" type="checkbox"/>	c1-deeplearning-tf-1-15-cu110-v20221026-debian-10	asia, eu, us	—	50 GB	Debian	tf-1-15-gpu-debian-10	—	⋮

Source: Google LLC

Excluir remove a imagem, e Depreciar marca a imagem como não mais suportada e permite que você especifique uma imagem de substituição para usar daqui para frente. Imagens depreciadas do Google estão disponíveis para uso, mas podem não ser corrigidas para falhas de segurança ou outras atualizações. A depreciação é uma maneira útil de informar aos usuários da imagem que ela não é mais suportada e que eles devem planejar testar suas aplicações com as versões mais novas e suportadas da imagem. Eventualmente, imagens depreciadas não estarão mais disponíveis, e os usuários das imagens depreciadas precisarão usar diferentes versões.

Depois de ter criado uma imagem, você pode criar uma instância usando essa imagem selecionando a opção Criar Instância na linha de comandos acima da listagem de imagens.

Gerenciando uma Única Instância de Máquina Virtual com Cloud Shell e a Linha de Comando

Além de gerenciar VMs através do console, você pode gerenciar recursos de computação usando a linha de comando. Os mesmos comandos podem ser usados no Cloud Shell ou no seu ambiente local depois de ter instalado o Google Cloud SDK, que foi abordado no Capítulo 5, “Computando com Máquinas Virtuais do Compute Engine”.

FIGURE 6.13 Cloud Console form for creating an image

[←](#) Create an image

Name * [?](#)

Name is permanent

Source * [?](#)

Source disk * [?](#)

Location [?](#)

Multi-regional
 Regional

Select location

Family [?](#)

Description

Labels [+ ADD LABEL](#)

Encryption

Data is encrypted automatically. Select an encryption key management solution.

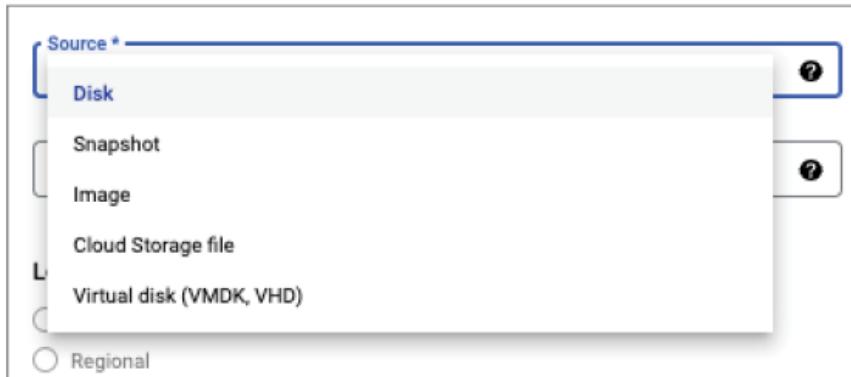
Google-managed encryption key
No configuration required

Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

Customer-supplied encryption key (CSEK)
Manage outside of Google Cloud

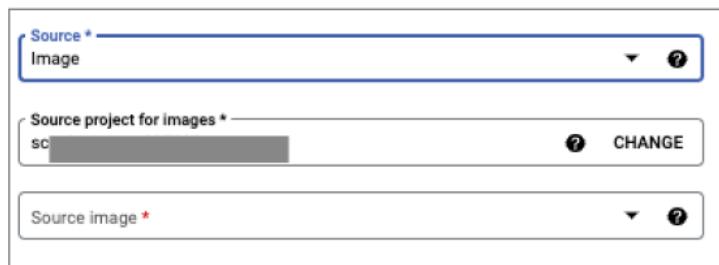
Source: Google LLC

FIGURE 6.14 Options for the source of an image



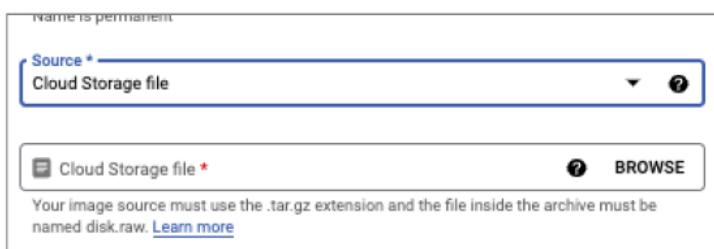
Source: Google LLC

FIGURE 6.15 When using an image as a source, you can choose a source image from another project.



Source: Google LLC

FIGURE 6.16 When using a Cloud Storage file as a source, you browse your storage buckets for a file.



Source: Google LLC

Esta seção descreve os comandos mais importantes para trabalhar com instâncias. Comandos têm seus próprios conjuntos específicos de parâmetros; no entanto, todos os comandos gcloud suportam conjuntos de flags. Estes são referidos como flags gcloud-wide, também conhecidos como flags globais gcloud, e incluem o seguinte:

- --account especifica uma conta do Google Cloud para usar, sobrescrevendo a conta padrão.
- --configuration usa um arquivo de configuração nomeado que contém pares chave-valor.

- --flatten gera registros chave-valor separados quando uma chave tem múltiplos valores.
- --format especifica um formato de saída, como padrão (legível por humanos) CSV, JSON, YAML, texto ou outras opções possíveis.
- --help exibe uma mensagem de ajuda detalhada.
- --project especifica um projeto do Google Cloud para usar, sobrescrevendo o projeto padrão.
- --quiet desativa prompts interativos e usa padrões.
- --verbosity especifica o nível de mensagens de saída detalhadas. As opções são debug, info, warning e error.

Ao longo desta seção, comandos podem ter um parâmetro opcional --zone. Assumimos que uma zona padrão foi definida quando você executou gcloud init.

Iniciando Instâncias

Para iniciar uma instância, use o comando gcloud, especificando que você está trabalhando com um serviço de computação e especificamente com instâncias. Você também precisa indicar que estará iniciando uma instância especificando start, seguido pelo nome de uma ou mais instâncias.

A sintaxe do comando é a seguinte:

```
gcloud compute instances start NOMES_DAS_INSTANCIAS
```

Um exemplo é o seguinte:

```
gcloud compute instances start instance-1 instance-2
```

O comando de iniciar instância também aceita parâmetros opcionais. O parâmetro --async retorna imediatamente sem esperar que as operações sejam concluídas. A opção -verbose em muitos comandos Linux fornece funcionalidade semelhante. Aqui está um exemplo:

```
gcloud compute instances start instance-1 instance-2 --async
```

O Google Cloud precisa saber em qual zona criar uma instância. Isso pode ser especificado com o parâmetro --zone da seguinte forma:

```
gcloud compute instances start ch06-instance-1 ch06-instance-2 --zone=us-central1-c
```

Você pode obter uma lista de zonas com o seguinte comando:

```
gcloud compute zones list
```

Se nenhuma zona for especificada, o comando solicitará uma.

Parando Instâncias

Para parar uma instância, use gcloud compute instances e especifique stop seguido pelo nome de uma ou mais instâncias.

A sintaxe do comando é a seguinte:

```
gcloud compute instances stop NOMES_DAS_INSTANCIAS
```

Aqui está um exemplo:

```
gcloud compute instances stop instance-3 instance-4
```

Assim como o comando de iniciar instância, o comando de parar aceita parâmetros opcionais:

```
gcloud compute instances stop ch06-instance-1 ch06-instance-2 --async
```

O Google Cloud precisa saber em qual zona está a instância a ser parada. Isso pode ser especificado com o parâmetro `--zone` da seguinte forma:

```
gcloud compute instances stop ch06-instance-1 ch06-instance-2 --zone=us-central1-c
```

Você pode obter uma lista de zonas com o seguinte comando:

```
gcloud compute zones list
```

Excluindo Instâncias

Quando você terminar de trabalhar com uma VM, você pode excluí-la com o comando `delete`.

Aqui está um exemplo:

```
gcloud compute instances delete instance-1
```

O comando `delete` aceita o parâmetro `--zone` para especificar onde a VM a ser excluída está localizada. Aqui está um exemplo:

```
gcloud compute instances delete ch06-instance-1 --zone=us-central1-b
```

Quando uma instância é excluída, os discos na VM podem ser excluídos ou salvos usando os parâmetros `--delete-disks` e `--keep-disks`, respectivamente. Você pode especificar `all` para manter todos os discos, `boot` para especificar a partição do sistema de arquivos raiz e `data` para especificar discos não de inicialização.

Por exemplo, o seguinte comando mantém todos os discos:

```
gcloud compute instances delete ch06-instance-1 --zone=us-central2-b --keep-disks=all
```

enquanto o seguinte exclui todos os discos não de inicialização:

```
gcloud compute instances delete ch06-instance-1 --zone=us-central2-b --delete-disks=data
```

Visualizando o Inventário de VMs

O comando para visualizar o conjunto de VMs em seu inventário é o seguinte:

```
gcloud compute instances list
```

Este comando aceita um nome opcional de uma instância. Para listar VMs em uma zona particular, você pode usar o seguinte:

```
gcloud compute instances list --filter="zone:ZONE"
```

onde ZONE é o nome de uma zona. Você pode listar várias zonas usando uma lista separada por vírgulas.

O parâmetro --limit é usado para limitar o número de VMs listadas, e o parâmetro --sort-by é usado para reordenar a lista de VMs especificando um campo de recurso. Você pode ver os campos de recurso para uma VM executando o seguinte:

```
gcloud compute instances describe
```

Trabalhando com Snapshots

Você pode criar um snapshot de um disco usando o seguinte comando:

```
gcloud compute disks snapshot DISK_NAME --snapshot-names=NAME
```

onde DISK_NAME é o nome de um disco e NAME é o nome do snapshot. Para visualizar uma lista de snapshots, use o seguinte:

```
gcloud compute snapshots list
```

Para informações detalhadas sobre um snapshot, use o seguinte:

```
gcloud compute snapshots describe NOME_DO_SNAPSHOT
```

onde NOME_DO_SNAPSHOT é o nome do snapshot a ser descrito. Para criar um disco, use isto:

```
gcloud compute disks create NOME_DO_DISCO --source-snapshot=NOME_DO_SNAPSHOT
```

Você também pode especificar o tamanho do disco e o tipo de disco usando os parâmetros --size e --type.

Aqui está um exemplo:

```
gcloud compute disks create disk-1 --source-snapshot=ch06-snapshot --size=100 --type=pd-standard
```

Isso criará um disco de 100 GB usando o ch06-snapshot e um disco persistente padrão.

Trabalhando com Imagens

O Google Cloud oferece uma ampla gama de imagens para usar ao criar uma VM; no entanto, você pode precisar criar uma imagem especializada por conta própria. Isso pode ser feito com o seguinte comando:

```
gcloud compute images create NOME_DA_IMAGEM
```

onde NOME_DA_IMAGEM é o nome dado às imagens. A fonte para as imagens é especificada usando um dos parâmetros de fonte, que são os seguintes:

- --source-disk
- --source-image
- --source-image-family
- --source-snapshot
- --source-uri

Os parâmetros source-disk, source-image e source-snapshot são usados para criar uma imagem usando um disco, imagem e snapshot, respectivamente. O parâmetro source-image-family usa a versão mais recente de uma imagem na família. Famílias são grupos de imagens relacionadas, que geralmente são diferentes versões da mesma imagem subjacente. O parâmetro source-uri permite especificar uma imagem usando um endereço web.

Uma imagem pode ter uma descrição e um conjunto de etiquetas. Estes são atribuídos usando os parâmetros --description e --labels.

Aqui está um exemplo de criação de uma nova imagem a partir de um disco:

```
gcloud compute images create image-1 --source-disk=disk-1
```

Você também pode excluir imagens quando elas não forem mais necessárias usando isto:

```
gcloud compute images delete NOME_DA_IMAGEM
```

Muitas vezes é útil armazenar imagens no Cloud Storage. Você pode exportar uma imagem para o Cloud Storage com o seguinte comando:

onde URI_DESTINO é o endereço de um bucket do Cloud Storage onde você deseja armazenar a imagem.

Introdução aos Grupos de Instâncias

Grupos de instâncias são conjuntos de VMs que são gerenciados como uma única entidade. Qualquer comando gcloud ou do console aplicado a um grupo de instâncias é aplicado a todos os membros do grupo de instâncias. O Google fornece dois tipos de grupos de instâncias: gerenciados e não gerenciados.

Grupos gerenciados consistem em grupos de VMs idênticas. Eles são criados usando um modelo de instância, que é uma especificação de uma configuração de VM, incluindo tipo de máquina, imagem de disco de inicialização, zona, etiquetas e outras propriedades de uma instância. Grupos de instâncias gerenciadas podem escalar automaticamente o número de instâncias em um grupo e ser usados com balanceamento de carga para distribuir cargas de trabalho pelo grupo de instâncias. Se uma instância em um grupo falhar, ela será recriada automaticamente. Grupos gerenciados são o tipo preferido de grupo de instâncias.

Grupos não gerenciados devem ser usados apenas quando você precisa trabalhar com configurações diferentes dentro de diferentes VMs no grupo.

Criando e Removendo Grupos de Instâncias e Modelos

Para criar um grupo de instâncias, você deve primeiro criar um modelo de grupo de instâncias. Para criar um modelo de instância, use o seguinte comando:

```
gcloud compute instance-templates create NOME_DO_MODELO
```

Você pode especificar uma VM existente como fonte do modelo de instância usando o parâmetro --source-instance. Aqui está um exemplo:

```
gcloud compute instance-templates create instance-template-1 --source-instance=instance-1
```

Modelos de grupo de instâncias também podem ser criados no console usando a página Template de Grupos de Instâncias, conforme mostrado na Figura 6.17.

Grupos de instâncias podem conter instâncias em uma única zona ou em uma região. O primeiro é chamado de grupo de instâncias gerenciadas zonal e o segundo é chamado de grupo de instâncias gerenciadas regional. Grupos de instâncias gerenciadas regionais são recomendados porque essa configuração distribui a carga de trabalho por zonas, aumentando a resiliência.

FIGURE 6.17 Instance group templates can be created in the console using a form similar to the create instance form.

Set up automatic management for a group of stateless VMs, including updates, regional deployments, load balancing, autoscaling, and autohealing. [Learn more](#)

Name * ?
Name is permanent

Description

Instance template *

Number of instances

Location

For higher availability, select multiple zones in a region instead of a single zone. [Learn more](#)

Single zone
 Multiple zones

Region * ? Zone * ?

Autoscaling

Use autoscaling to automatically add and remove instances to the group for periods of high and low load. [Learn more](#)

Autoscaling mode ?

Minimum number of instances * ? Maximum number of instances * ?

Autoscaling metrics

Use metrics to help determine when to scale the group. [Learn more](#)

CPU utilization: 60% (default)
Predictive autoscaling is off

[ADD METRIC](#)

Autoscaling schedules

Source: Google LLC

Você pode especificar uma política de distribuição. A distribuição uniforme distribuirá de maneira uniforme pelas zonas. A distribuição equilibrada distribuirá o mais uniformemente possível pelas zonas, com base nos recursos disponíveis. A distribuição "Qualquer" (Any) implantará instâncias gerenciadas nas zonas com base na disponibilidade e nas reservas.

Você pode remover modelos de instâncias excluindo-os da página de Modelo de Grupo de Instâncias no console. Selecione o modelo de grupo de instâncias marcando a caixa de seleção na lista de modelos e depois exclua-o clicando no ícone de excluir.

Você também pode excluir um modelo de grupo de instâncias usando o seguinte comando:

```
gcloud compute instance-templates delete NOME_DO_MODELO_DE_INSTANCIA
```

onde NOME_DO_MODELO_DE_INSTANCIA é o nome do modelo que você deseja excluir.

Para listar modelos e grupos de instâncias, use o seguinte:

```
gcloud compute instance-templates list
```

```
gcloud compute instance-groups managed list-instances
```

Para listar as instâncias em um grupo de instâncias, use o seguinte:

```
gcloud compute instance-groups managed list-instances  
NOME_DO_GRUPO_DE_INSTANCIAS
```

Balanceamento de Carga e Dimensionamento Automático de Grupos de Instâncias

Para implantar uma aplicação escalável e altamente disponível, você pode executar essa aplicação em um conjunto de instâncias com balanceamento de carga. O Google Cloud oferece vários tipos de balanceamento de carga, e todos exigem o uso de um grupo de instâncias.

Além do balanceamento de carga, os grupos de instâncias gerenciados podem ser configurados para dimensionamento automático. Você pode configurar uma política de dimensionamento automático para acionar a adição ou remoção de instâncias com base na utilização da CPU, métrica de monitoramento, capacidade de balanceamento de carga ou cargas de trabalho baseadas em fila.

Não Há Mais Planejamento de Capacidade Máxima

Antes do advento da nuvem, organizações de TI muitas vezes tinham que planejar suas compras de hardware em torno da carga máxima esperada. Isso é chamado de planejamento de capacidade máxima. Se há pouca variação na carga, o planejamento de capacidade máxima é uma abordagem sólida. Negócios com cargas de trabalho altamente variáveis, como varejistas nos Estados Unidos que têm alta demanda durante os últimos dois meses do ano, teriam que suportar capacidade ociosa por meses do ano. A computação em nuvem e o dimensionamento automático eliminaram a necessidade de planejamento de capacidade máxima. Servidores adicionais são adquiridos em minutos, não semanas ou meses. Quando a capacidade não é necessária, ela é reduzida. Grupos de instâncias automatizam o processo de adicionar e remover VMs, permitindo que engenheiros de nuvem ajustem quando adicionar e quando remover VMs.

Ao dimensionar automaticamente, certifique-se de deixar tempo suficiente para as VMs inicializarem ou desligarem antes de acionar outra mudança na configuração do cluster. Se o tempo entre verificações for muito curto, você pode descobrir que uma VM recentemente adicionada não foi totalmente iniciada antes que outra seja adicionada. Isso pode levar à adição de mais VMs do que as realmente necessárias.

Diretrizes para Gerenciamento de Máquinas Virtuais

Aqui estão algumas diretrizes para o gerenciamento de VMs:

- Use etiquetas e descrições. Isso ajudará você a identificar o propósito de uma instância e ajudar ao filtrar listas de instâncias.
- Use grupos de instâncias gerenciados para habilitar o dimensionamento automático e o balanceamento de carga. Estes são chaves para implantar serviços escaláveis e altamente disponíveis.
- Use GPUs para processamento intensivo numérico, como aprendizado de máquina e computação de alto desempenho. Para algumas aplicações, GPUs podem oferecer um benefício de desempenho maior do que adicionar outra CPU.
- Use snapshots para salvar o estado de um disco ou para fazer cópias. Estes podem ser salvos no Cloud Storage e agir como backups.
- Use instâncias preemptíveis para cargas de trabalho que podem tolerar interrupções. VMs spot são VMs de custo mais baixo que são adequadas por 60% a 91% menos do que VMs com preços padrão. Elas são instâncias preemptíveis mas não são limitadas a tempos de execução máximos de 24 horas como as VMs preemptíveis originais.

Resumo

Neste capítulo, você aprendeu como gerenciar instâncias únicas de VM e grupos de instâncias. Instâncias únicas de VM podem ser criadas, configuradas, paradas, iniciadas e excluídas usando o Cloud Console ou comandos gcloud do Cloud Shell ou da sua máquina local, se você tiver o SDK instalado.

Snapshots são cópias de discos e são úteis como backups e para copiar dados para outras instâncias. Imagens são backups completos de um disco de inicialização, então são usadas para criar VMs. Snapshots feitos a partir de um disco de inicialização também podem ser usados para criar uma VM.

O principal comando usado para gerenciar VMs é o comando gcloud compute instances. O gcloud usa uma estrutura hierárquica para ordenar os elementos do comando. O comando começa com gcloud, seguido por um componente do Google Cloud, como compute para o Compute Engine, seguido por um tipo de entidade, como instâncias ou snapshots. Uma operação é então especificada, como criar, excluir, listar ou descrever.

GPUs podem ser anexadas a instâncias que têm bibliotecas de GPU instaladas no sistema operacional. GPUs são usadas para tarefas intensivas de computação, como construir modelos de aprendizado de máquina.

Grupos de instâncias são grupos de instâncias que são gerenciados juntos. Grupos de instâncias gerenciadas têm instâncias que são iguais. Esses grupos suportam balanceamento de carga e dimensionamento automático.

Essenciais para o Exame

Entenda como navegar no Cloud Console. O Cloud Console é a interface gráfica para trabalhar com o Google Cloud. Você pode criar, configurar, excluir e listar instâncias de VM na área do Compute Engine do console.

Entenda como instalar o Cloud SDK. O Cloud SDK permite configurar variáveis de ambiente padrão, como uma zona preferencial, e emitir comandos a partir da linha de comando. Se você usar o Cloud Shell, o Cloud SDK já estará instalado.

Saiba como criar uma VM no console e na linha de comando. Você pode especificar o tipo de máquina, escolher uma imagem e configurar discos com o console. Você pode usar comandos na linha de comando para listar e descrever, e você pode encontrar as mesmas informações no console. Entenda quando usar imagens personalizadas e como depreciá-las. Imagens são cópias dos conteúdos de um disco e são usadas para criar VMs. Depreciar marca uma imagem como não mais suportada.

Entenda por que GPUs são usadas e como anexá-las a uma VM. GPUs são usadas para operações intensivas de computação; um caso de uso comum para o uso de GPUs é aprendizado de máquina. É melhor usar uma imagem que tenha bibliotecas de GPU instaladas. Entenda como determinar quais locais têm GPUs disponíveis, pois existem algumas restrições. A CPU deve ser compatível com a GPU selecionada, e GPUs não podem ser anexadas a máquinas de memória compartilhada. Conheça como os custos de GPU são cobrados.

Entenda imagens e snapshots. Snapshots salvam o conteúdo dos discos para fins de backup e compartilhamento de dados. Imagens salvam o sistema operacional e configurações relacionadas para que você possa criar cópias idênticas da instância.

Entenda grupos de instâncias e modelos de grupos de instâncias. Grupos de instâncias são conjuntos de instâncias gerenciadas como uma única entidade. Modelos de grupo de instâncias especificam a configuração de um grupo de instâncias e as instâncias nele. Grupos de instâncias gerenciados suportam dimensionamento automático e balanceamento de carga.

Questões para Revisão

Você pode encontrar as respostas no Apêndice.

1. Qual página no Google Cloud Console você usaria para criar uma única instância de uma VM?
 - A. Compute Engine
 - B. App Engine
 - C. Kubernetes Engine
 - D. Cloud Functions
2. Você visualiza uma lista de instâncias de VMs Linux no console. Todas têm endereços IP públicos atribuídos. Você percebe que a opção SSH está desabilitada para uma das instâncias. Por que isso pode ser o caso?
 - A. A instância é preemptível e, portanto, não suporta SSH.
 - B. A instância está parada.
 - C. A instância foi configurada com a opção No SSH.
 - D. A opção SSH nunca é desabilitada.
3. Você notou um tempo de resposta incomumente lento ao emitir comandos para um servidor Linux e decide reiniciar a máquina. Qual comando você usaria no console para reiniciar?
 - A. Reboot
 - B. Reset
 - C. Restart
 - D. Shutdown seguido por Startup
4. No console, você pode filtrar a lista de instâncias de VM por qual dos seguintes?
 - A. Apenas etiquetas
 - B. Apenas membros do grupo de instâncias gerenciadas
 - C. Etiquetas, status ou prevenção de exclusão
 - D. Apenas etiquetas e status
5. Você estará construindo vários modelos de aprendizado de máquina em uma instância e anexando GPU à instância. Quando você executa seus modelos de aprendizado de máquina, eles levam um tempo anormalmente longo para rodar. Parece que a GPU não está sendo usada. Qual pode ser a causa disso?
 - A. As bibliotecas de GPU não estão instaladas.
 - B. O sistema operacional é baseado em Ubuntu.

- C. Você não tem pelo menos oito CPUs na instância.
 - D. Não há espaço suficiente em disco persistente disponível.
6. Ao adicionar uma GPU a uma instância, você deve garantir que:
- A. As escolhas de GPU e CPU sejam compatíveis.
 - B. A instância seja preemptível.
 - C. A instância não tenha discos não inicializáveis anexados.
 - D. A instância esteja executando Ubuntu 18.02 ou posterior.
7. Você está usando snapshots para salvar cópias de um disco de 100 GB. Você faz um snapshot e depois adiciona 10 GB de dados. Você cria um segundo snapshot. Quanto espaço de armazenamento é usado no total para os dois snapshots (assumindo que não há compressão)?
- A. 210 GB, com 100 GB para o primeiro e 110 GB para o segundo
 - B. 110 GB, com 100 GB para o primeiro e 10 GB para o segundo
 - C. 110 GB, com 110 GB para o segundo (o primeiro snapshot é automaticamente deletado)
 - D. 221 GB, com 100 GB para o primeiro, 110 GB para o segundo, mais 10 por cento do segundo snapshot (11 GB) para sobrecarga de metadados
8. Você decidiu delegar a tarefa de fazer snapshots de backup a um membro de sua equipe. Qual papel você precisaria conceder ao seu membro da equipe para criar snapshots?
- A. Administrador de Imagem de Computação
 - B. Administrador de Armazenamento
 - C. Administrador de Snapshot de Computação
 - D. Administrador de Armazenamento de Computação
9. A fonte de uma imagem pode ser:
- A. Apenas discos
 - B. Apenas snapshots ou discos
 - C. Discos, snapshots ou outra imagem
 - D. Discos, snapshots ou qualquer arquivo de exportação de banco de dados
10. Você construiu imagens usando o Ubuntu 18.04 e agora quer que os usuários começem a usar o Ubuntu 20.04. Você não quer apenas deletar as imagens baseadas no Ubuntu 18.04, mas quer que os usuários saibam que eles devem começar a usar o Ubuntu 20.04. Qual recurso das imagens você usaria para realizar isso?

- A. Redirecionamento
 - B. Depreciado
 - C. Não suportado
 - D. Migração
11. Você quer gerar uma lista de VMs em seu inventário e ter os resultados em formato JSON. Que comando você usaria?
- A. gcloud compute instances list
 - B. gcloud compute instances describe
 - C. gcloud compute instances list --format=json
 - D. gcloud compute instances list --output=json
12. Você gostaria de entender detalhes de como o Google Cloud inicia uma instância virtual. Qual parâmetro opcional você usaria ao iniciar uma instância para exibir esses detalhes?
- A. --verbose
 - B. --async
 - C. --describe
 - D. --details
13. Qual comando deletará uma instância chamada ch06-instance-3?
- A. gcloud compute instances delete instance=ch06-instance-3
 - B. gcloud compute instance stop ch06-instance-3
 - C. gcloud compute instances delete ch06-instance-3
 - D. gcloud compute delete ch06-instance-3
14. Você está prestes a excluir uma instância chamada ch06-instance-1, mas quer manter seu disco de inicialização. Você não quer manter outros discos anexados. Qual comando gcloud você usaria?
- A. gcloud compute instances delete ch06-instance-1 --keep-disks=boot
 - B. gcloud compute instances delete ch06-instance-1 --save-disks=boot
 - C. gcloud compute instances delete ch06-instance-1 --keep-disks=filesystem
 - D. gcloud compute delete ch06-instance-1 --keep-disks=filesystem
15. Você quer visualizar uma lista de campos que você pode usar para ordenar uma lista de instâncias. Que comando você usaria para ver os nomes dos campos?
- A. gcloud compute instances list
 - B. gcloud compute instances describe

- C. gcloud compute instances list --detailed
 - D. gcloud compute instances describe --detailed
16. Você está implantando uma aplicação que precisará escalar e ser altamente disponível. Qual desses componentes do Compute Engine ajudará a alcançar escalabilidade e alta disponibilidade?
- A. Instâncias preemptíveis
 - B. Grupos de instâncias
 - C. Cloud Storage
 - D. GPUs
17. Antes de criar um grupo de instâncias, o que você precisa criar?
- A. Instâncias no grupo de instâncias
 - B. Modelo de instância
 - C. Imagem de disco de inicialização
 - D. Snapshot de origem
18. Como você deletaria um grupo de instâncias usando a linha de comando?
- A. gcloud compute instances instance-template delete
 - B. gcloud compute instance-templates delete
 - C. gcloud compute delete instance-template
 - D. gcloud compute delete instance-templates
19. O que pode ser a base para dimensionar um grupo de instâncias?
- A. Utilização da CPU e atualizações do sistema operacional
 - B. Uso do disco e utilização da CPU apenas
 - C. Latência de rede, capacidade de balanceamento de carga e utilização da CPU
 - D. Uso do disco e atualizações do sistema operacional apenas
20. Um arquiteto está movendo uma aplicação legada para o Google Cloud e quer minimizar as mudanças na arquitetura existente enquanto administra o cluster como uma única entidade. A aplicação legada é executada em um cluster com平衡amento de carga que roda nós com duas configurações diferentes. As duas configurações são necessárias devido a decisões de design feitas vários anos atrás. A carga na aplicação é consistente, então raramente há necessidade de escalar para cima ou para baixo. Que recurso do Google Cloud Compute Engine você recomendaria usar?
- A. Instâncias preemptíveis
 - B. Grupos de instâncias não gerenciados

C. Grupos de instâncias gerenciados

D. GPUs

Capítulo 7

Computação com Kubernetes

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO DE ENGENHEIRO DE NUVEM ASSOCIADO DO GOOGLE:

✓✓ 3.2 Implantando e implementando recursos do Google Kubernetes Engine

Este capítulo apresenta o Kubernetes, um sistema de orquestração de contêineres criado e disponibilizado em código aberto pelo Google. Você aprenderá sobre a arquitetura do Kubernetes e as maneiras como ele gerencia cargas de trabalho em nós de um cluster. Você também aprenderá como gerenciar recursos do Kubernetes com o Cloud Console, Cloud Shell e Cloud SDK. O capítulo também aborda como implantar pods de aplicativos (uma estrutura do Kubernetes) e monitorar e registrar recursos do Kubernetes.

Introdução ao Kubernetes Engine

O Kubernetes Engine é um serviço gerenciado do Kubernetes no Google Cloud. Com esse serviço, os clientes do Google Cloud podem criar e manter seus próprios clusters do Kubernetes sem ter que gerenciar a plataforma Kubernetes. O Google Kubernetes Engine às vezes é abreviado como GKE.

O Kubernetes executa contêineres em um cluster de máquinas virtuais (VMs). Ele determina onde executar contêineres, monitora a saúde dos contêineres e gerencia todo o ciclo de vida das instâncias de VM. Esse conjunto de tarefas é conhecido como orquestração de contêineres.

Pode parecer que um cluster do Kubernetes é como um grupo de instâncias, que foi discutido no Capítulo 6, "Gerenciando Máquinas Virtuais". Existem algumas semelhanças, e, de fato, o GKE usa grupos de instâncias para gerenciar as VMs subjacentes em um cluster GKE.

Contêineres oferecem um meio altamente portátil e leve de distribuir e escalar suas aplicações ou cargas de trabalho, como VMs, sem replicar o sistema operacional convidado. Eles podem iniciar e parar muito mais rápido (geralmente em segundos) e usar menos recursos. Você pode pensar em um contêiner como semelhante a contêineres de transporte para aplicações e cargas de trabalho. Como contêineres de transporte que podem viajar em navios, trens e caminhões sem reconfiguração, contêineres de aplicativos podem ser movidos de laptops de desenvolvimento para servidores de teste e produção sem reconfiguração.

Grupos de instâncias têm monitoramento configurável e podem reiniciar instâncias que falham, mas o Kubernetes tem muito mais flexibilidade com relação à manutenção de um cluster de servidores.

Vamos olhar para a arquitetura do Kubernetes, que consiste em vários objetos e um conjunto de controladores.

Tenha em mente que, ao usar o Kubernetes Engine, você gerenciará o Kubernetes e suas aplicações e cargas de trabalho executando em contêineres na plataforma Kubernetes.

Arquitetura do Cluster Kubernetes

Um cluster Kubernetes consiste em um plano de controle e uma ou mais máquinas trabalhadoras chamadas nós. O plano de controle gerencia o cluster e pode ser replicado e distribuído para alta disponibilidade e tolerância a falhas.

O plano de controle gerencia serviços fornecidos pelo Kubernetes, como a API do Kubernetes, controladores e agendadores. Todas as interações com o cluster são feitas através do plano de controle usando a API do Kubernetes. O plano de controle emite o comando que executa uma ação em um nó. Os usuários também podem interagir com um cluster usando o comando kubectl.

Os componentes básicos do Kubernetes são:

- API Server, que é um componente do plano de controle que expõe a API do Kubernetes
- Scheduler, um componente do plano de controle que atribui pods aos nós
- Controller Manager, um componente do plano de controle que gerencia controladores de recursos, tais como controlador de nó, controlador de trabalho e controlador de conta de serviço
- etcd, um armazenamento de chave-valor altamente disponível
- Kubelet, um agente que roda em cada nó em um cluster
- Container Runtime, o software responsável por executar contêineres
- Kube-proxy, um proxy de rede que roda em cada nó no cluster

Os nós executam as cargas de trabalho executadas no cluster. Os nós são VMs que executam contêineres configurados para executar uma aplicação. Os nós são primariamente controlados pelo plano de controle, mas alguns comandos podem ser executados manualmente. Os nós executam um agente chamado kubelet, que é o serviço que se comunica com o plano de controle.

Quando você cria um cluster GKE, pode especificar um tipo de máquina. Essas VMs executam sistemas operacionais especializados otimizados para executar contêineres. Parte da memória e da CPU é reservada para o Kubernetes e, portanto, não está disponível para aplicações executadas no nó.

O Kubernetes organiza o processamento em cargas de trabalho. Existem vários objetos organizadores que compõem a funcionalidade principal de como o Kubernetes processa cargas de trabalho.

Objetos Kubernetes

As cargas de trabalho são distribuídas entre os nós em um cluster Kubernetes. Para entender como o trabalho é distribuído, é importante entender alguns conceitos básicos, em particular os seguintes:

- Pods
- Services
- Deployments
- ReplicaSets
- StatefulSets
- Job
- Volumes
- Namespaces
- Pools de nós

Cada um desses objetos contribui para a organização lógica das cargas de trabalho.

Pods

Pods são instâncias únicas de um processo em execução em um cluster. Pods contêm pelo menos um contêiner. Eles frequentemente executam um único contêiner, mas podem executar vários contêineres. Contêineres múltiplos são usados quando dois ou mais contêineres precisam compartilhar recursos ou estão intimamente acoplados. Pods também utilizam redes e armazenamentos compartilhados entre contêineres. Cada pod recebe um endereço IP único e um conjunto de portas. Contêineres se conectam a uma porta. Vários contêineres em um pod se conectam a portas diferentes e podem se comunicar entre si em localhost. Esta estrutura é projetada para suportar a execução de uma instância de uma aplicação dentro do cluster como um pod. Um pod permite que seus contêineres se comportem como se estivessem rodando em uma VM isolada, compartilhando armazenamento comum, um endereço IP e um conjunto de portas. Fazendo isso, você pode implantar várias instâncias da mesma aplicação, ou diferentes instâncias de diferentes aplicações no mesmo nó ou em nós diferentes, sem ter que mudar sua configuração.

Pods tratam os múltiplos contêineres como uma única entidade para fins de gerenciamento.

Pods são geralmente criados em grupos. Rélicas são cópias de pods e constituem um grupo de pods que são gerenciados como uma unidade. Pods suportam escalonamento automático também. Pods são considerados efêmeros; ou seja, espera-se que eles terminem. Se um pod estiver doente — por exemplo, se estiver preso em um modo de espera ou travando repetidamente — ele é terminado. O mecanismo que gerencia o escalonamento e o monitoramento de saúde é conhecido como controlador.

Serviços

Já que os pods são efêmeros e podem ser terminados por um controlador, outros serviços que dependem de pods não devem ser fortemente acoplados a pods particulares. Por exemplo, mesmo que os pods tenham endereços IP únicos, aplicações não devem depender desse endereço IP para alcançar uma aplicação. Se o pod com esse endereço for terminado e outro for criado, ele pode ter outro endereço IP. O endereço IP pode ser reatribuído a outro pod executando um contêiner diferente.

O Kubernetes fornece um nível de indireção entre aplicações rodando em pods e outras aplicações que os chamam: isso é chamado de serviço. Um serviço, na terminologia do Kubernetes, é um objeto que fornece pontos de acesso API com um endereço IP estável que permite aplicações descobrirem pods rodando uma aplicação particular. Serviços se atualizam quando mudanças são feitas em pods, então eles mantêm uma lista atualizada de pods rodando uma aplicação.

Implantações

Outro conceito importante no Kubernetes é a implantação (deployment). Implantações são conjuntos de pods idênticos. Os membros do conjunto podem mudar conforme alguns pods são terminados e outros são iniciados, mas todos estão rodando a

mesma aplicação. Os pods todos rodam a mesma aplicação porque são criados usando o mesmo modelo de pod.

Um modelo de pod é uma definição de como rodar um pod. A descrição de como definir o pod é uma especificação de pod. O Kubernetes usa essa definição para manter um pod no estado especificado no modelo. Ou seja, se a especificação tem um número mínimo de pods que deveriam estar na implantação e o número cai abaixo disso, então pods adicionais serão adicionados à implantação chamando um ReplicaSet.

ReplicaSets

Um ReplicaSet é um controlador usado por uma implantação que garante o número correto de pods idênticos rodando. Por exemplo, se um pod é determinado como não saudável, um controlador terminará esse pod. O ReplicaSet detectará que não há pods suficientes para aquela aplicação ou carga de trabalho rodando e criará outro. ReplicaSets também são usados para atualizar e deletar pods. Em geral, é uma boa prática usar implantação em vez de ReplicaSets, a menos que você requeira orquestração de atualização personalizada ou não requeira atualizações de forma alguma.

StatefulSets

Implantações são bem adequadas para aplicações sem estado. Essas são aplicações que não precisam manter um registro do seu estado. Por exemplo, uma aplicação que chama uma API para realizar um cálculo sobre os valores de entrada não precisa manter um registro das chamadas ou cálculos anteriores. Uma aplicação que chama essa API pode alcançar um pod diferente a cada vez que faz uma chamada. No entanto, há momentos em que é vantajoso ter um único pod respondendo a todas as chamadas de um cliente durante uma única sessão.

StatefulSets são como implantações, mas eles atribuem identificadores únicos aos pods. Isso permite que o Kubernetes rastreie qual pod é usado por qual cliente e os mantenha juntos. StatefulSets são usados quando uma aplicação precisa de um identificador de rede único ou armazenamento persistente estável.

Trabalhos (Jobs)

Um trabalho é uma abstração sobre uma carga de trabalho. Trabalhos criam pods e os executam até que a aplicação complete uma carga de trabalho. As especificações de trabalho são especificadas em um arquivo de configuração e incluem especificações sobre o contêiner a ser usado e qual comando executar.

Volumes

Volumes são um mecanismo de armazenamento fornecido pelo Kubernetes. Volumes armazenam dados independentemente da vida de um pod. Se um pod falha e é reiniciado, o conteúdo de um volume anexado ao pod falhado continuará a existir após o pod ser reiniciado, e esse volume será anexado à nova instância do pod. Isso garante que, se um pod travar ou for reiniciado, os dados salvos em um volume estarão disponíveis para o pod de substituição. Volumes também são usados para compartilhar arquivos entre contêineres executando em um pod.

Namespaces

Um namespace é uma abstração lógica para separar grupos de recursos em um cluster. Namespaces são usados quando clusters hospedam uma variedade de projetos, equipes ou outros grupos que podem ter políticas ou requisitos diferentes para usar os recursos do cluster. O Kubernetes cria um namespace padrão, que é usado para objetos sem outro namespace definido. O Kubernetes também cria namespaces para administração do cluster.

Pools de Nós

Um pool de nós é um conjunto de nós em um cluster que têm a mesma configuração. Quando o cluster é criado pela primeira vez, todos os nós estão no mesmo pool de nós. Você pode adicionar outros nós e pools de nós após o cluster ser criado. Pools de nós são úteis se você quiser agrupar nós com características semelhantes, como nós que executam em máquinas virtuais preemptíveis. Um pool de nós de VMs preemptíveis permitiria atribuir algumas cargas de trabalho a nós nessas VMs preemptíveis, evitando que outras cargas de trabalho sejam executadas nelas.

Agora que você está familiarizado com como o Kubernetes é organizado e como as cargas de trabalho são executadas, vamos cobrir como implantar um cluster Kubernetes usando o Kubernetes Engine.

Implantando Clusters Kubernetes

Clusters Kubernetes podem ser implantados usando o Cloud Console ou a linha de comando no Cloud Shell, ou seu ambiente local se o Cloud SDK estiver instalado.

Implantando Clusters Kubernetes Usando o Cloud Console

Para usar o Kubernetes Engine, você precisará habilitar a API do Kubernetes Engine. Uma vez que a API esteja habilitada, você pode navegar até a página do Kubernetes Engine no Cloud Console. A Figura 7.1 mostra a página de visão geral.

Ao criar um cluster, você terá a opção de criar o cluster no modo padrão ou no modo piloto automático. No modo padrão, você paga pelos recursos do cluster que provisiona, gerencia a infraestrutura de nós e determina a configuração dos nós. No modo piloto automático, o GKE gerencia a infraestrutura do cluster e dos nós, e você paga apenas pelos recursos usados quando suas aplicações estão em execução. Clusters no modo piloto automático usam configurações de cluster pré-configuradas e otimizadas (veja a Figura 7.2). O modo piloto automático é o modo recomendado para usar o GKE.

FIGURE 7.1 The Overview page of the Kubernetes Engine section of Cloud Console

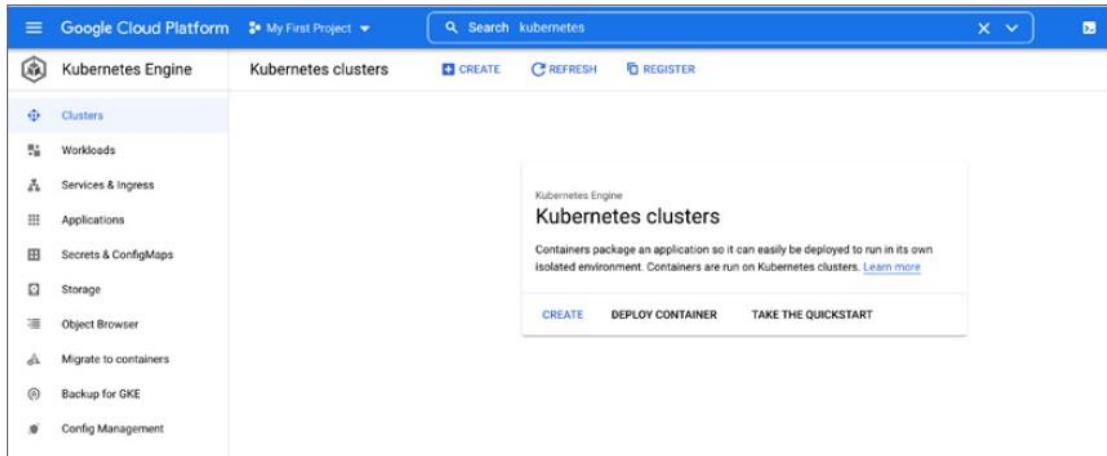
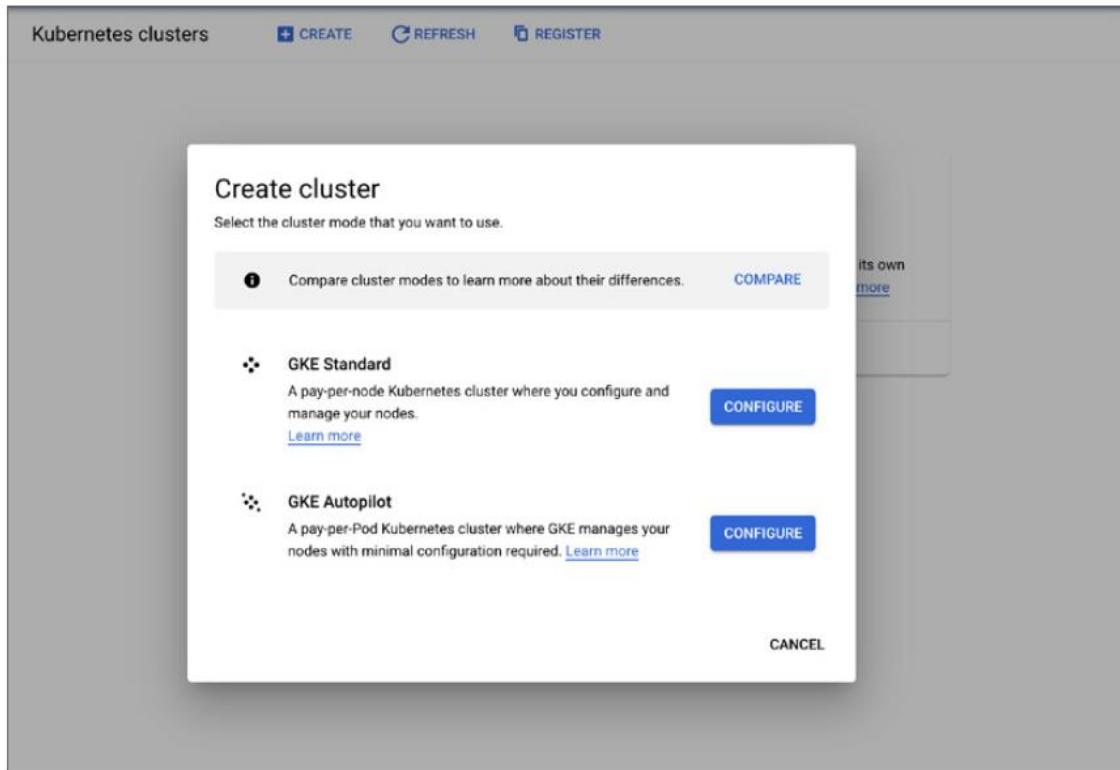
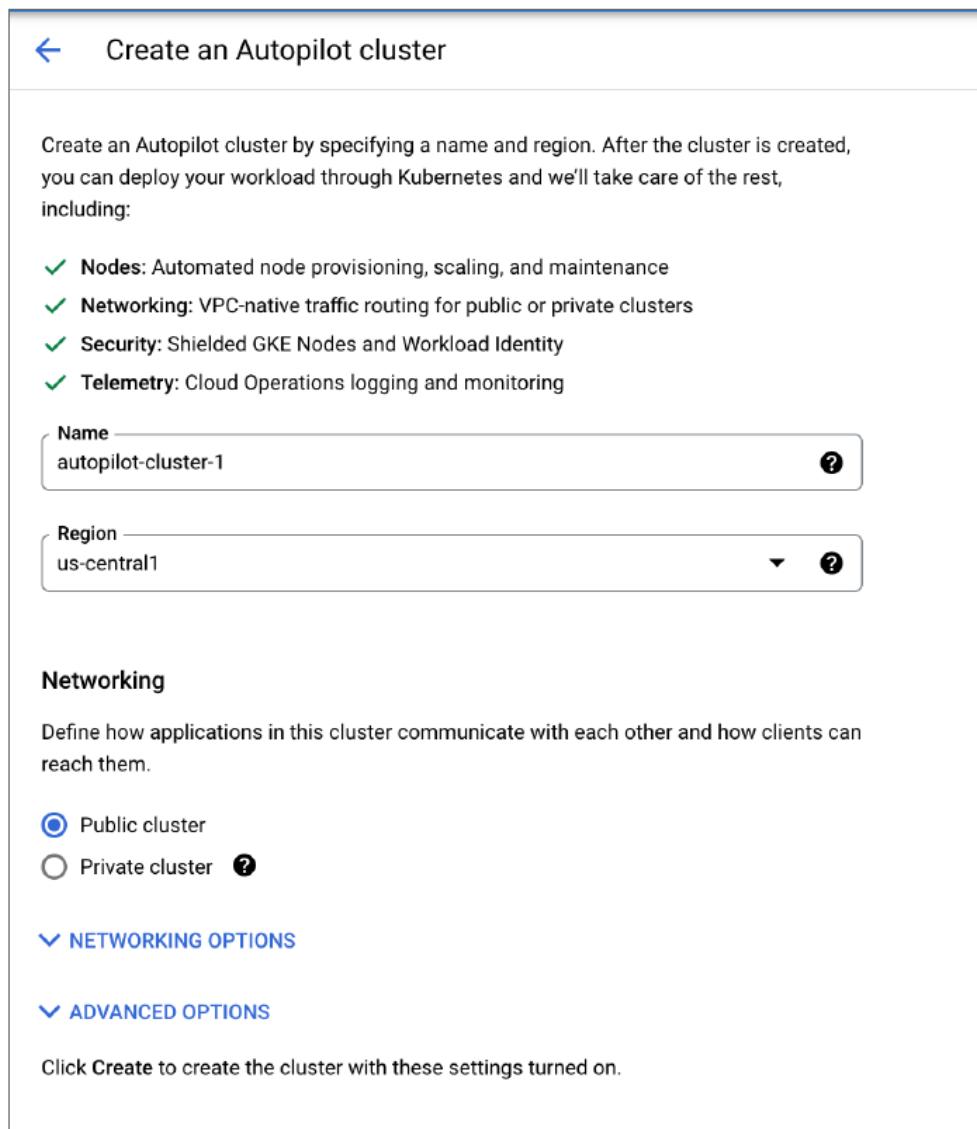


FIGURE 7.2 When creating a GKE, you specify standard mode or autopilot mode.



Quando você clica para criar um cluster no modo piloto automático, o GKE gerenciará e configurará automaticamente a infraestrutura de nós, roteamento de tráfego nativo da VPC para clusters públicos e privados, usará nós GKE Blindados, bem como logs e monitoramento. Você especificará um nome para o cluster, descrição do cluster e uma região. Você também especifica se o cluster é privado ou público. Em clusters privados, os nós têm apenas endereços IP privados e toda a comunicação entre o plano de controle e os nós é feita apenas por endereços privados. Veja a Figura 7.3.

FIGURE 7.3 Creating an autopilot GKE cluster



Expandindo a área de Opções de Rede na página Criar um Cluster no Modo Piloto Automático mostra configurações de rede adicionais, como você pode ver na Figura 7.4. Você pode habilitar redes autorizadas pelo plano de controle para bloquear endereços IP de origem não confiáveis e não pertencentes ao Google Cloud de acessar o plano de controle usando HTTPS. Você também pode especificar uma rede, sub-rede de nós e faixas de endereços para pods e serviços. Ao especificar faixas de endereços, você usa a notação CIDR; por exemplo, 192.168.0.0/16.

FIGURE 7.4 Networking options in autopilot mode

The screenshot shows the 'Networking' configuration section. At the top, it says 'Define how applications in this cluster communicate with each other and how clients can reach them.' There are two radio button options: 'Public cluster' (selected) and 'Private cluster'. Below this is a checkbox for 'Enable control plane authorized networks' which is unchecked. There are four input fields with dropdown menus: 'Network *' set to 'default', 'Node subnet *' set to 'default', 'Cluster default pod address range' set to '/17' with an example '192.168.0.0/16', and 'Service address range' set to '/22' with an example '192.168.0.0/16'. Each input field has a question mark icon in its corner.

Na área de Opções Avançadas na página Criar um Cluster no Modo Piloto Automático, você pode especificar uma janela de manutenção para definir um tempo para executar operações de manutenção rotineiras do Kubernetes. Por padrão, essas operações podem ser executadas a qualquer momento. Você também pode habilitar recursos de segurança, incluindo Google Groups para RBAC, para conceder papéis aos membros de um Grupo do Google Workspace, criptografia de segredos na camada de aplicação para criptografar segredos armazenados no etcd (parte do plano de controle), e habilitar o uso de uma chave gerenciada pelo cliente para criptografar o disco de boot dos nós. Você pode adicionar etiquetas e uma descrição ao cluster. Veja a Figura 7.5.

Da listagem de clusters, você pode editar, excluir e conectar-se a um cluster. Quando você clica em Conectar, recebe um comando gcloud para conectar ao cluster a partir da linha de comando. Você também tem a opção de visualizar a página de Cargas de Trabalho, como mostrado na Figura 7.6.

Quando você escolhe configurar um cluster no modo padrão usando o console cloud, você verá um formulário como o mostrado na Figura 7.7. Você especificará um nome e localização do cluster. Se você escolher criar um cluster zonal, a localização será uma zona. Se você escolher criar um cluster regional, a localização será uma região. Clusters regionais, por padrão, têm nós em três zonas, mas você pode especificar localizações de nós padrão se quiser especificar zonas específicas para executar nós.

Por padrão, os clusters são criados com uma configuração de canal de lançamento, que habilita a atualização automática do software do cluster. Se você deseja mais controle

sobre o processo de atualização, pode optar por configurar um canal estático. Veja a Figura 7.7.

FIGURE 7.5 Advanced options in autopilot mode

The screenshot shows a configuration interface for an Autopilot cluster. It is divided into several sections:

- Automation**: Contains a checkbox for "Enable Maintenance Window" with a help icon and a note: "* Indicates required field". Below it is a button labeled "+ ADD MAINTENANCE EXCLUSION".
- Security**: Contains three checkboxes: "Enable Google Groups for RBAC", "Enable Application-layer Secrets Encryption", and "Enable customer-managed encryption for boot disk", each with a help icon.
- Metadata**: A section for adding a description and labels. It includes a "Description" input field with a help icon and a "Labels" section.
- Labels**: A section for organizing resources using key/value pairs. It includes a "Learn more" link and a button "+ ADD LABEL".
- HIDE ADVANCED OPTIONS**: A button at the bottom left of the configuration area.

FIGURE 7.6 Once the autopilot clusters are deployed, it will be listed on the GKE page of the console.

The screenshot shows the GKE console's "Clusters" page. The sidebar on the left lists "Clusters", "Workloads", "Services & Ingress", "Applications", and "Secrets & ConfigMaps". The main area displays the "OVERVIEW" tab for a cluster named "autopilot-cluster-1". The cluster details are as follows:

Status	Name	Location	Mode	Number of nodes	Total vCPUs	Total memory	Notifications	Labels
Green checkmark	autopilot-cluster-1	us-central1	Autopilot	0	0	0 GB	-	-

FIGURE 7.7 Initial steps to configure a standard cluster

Cluster basics

The new cluster will be created with the name, version, and in the location you specify here. After the cluster is created, name and location can't be changed.

Tip To experiment with an affordable cluster, try [My first cluster in the Cluster set-up guides](#)

Name cluster-1 [?](#)

Location type
 Zonal
 Regional

Zone us-central1-c [?](#)

Specify default node locations [?](#)
Current default: us-central1-c

Control plane version
Choose a release channel for automatic management of your cluster's version and upgrade cadence. Choose a static version for more direct management of your cluster's version. [Learn more](#).

Static version
 Release channel

Release channel Regular channel (default) [▼](#)

Version 1.21.9-gke.1002 (default) [▼](#)

Como outros serviços do Google Cloud, o Kubernetes Engine pode ser gerenciado usando a linha de comando. O comando básico para trabalhar com o Kubernetes Engine é o seguinte comando gcloud:

```
gcloud container
```

Este comando gcloud tem muitos parâmetros, incluindo os seguintes:

- Projeto
- Zona
- Tipo de máquina
- Tipo de imagem
- Tipo de disco

- Tamanho do disco
- Número de nós

Um comando básico para criar um cluster no modo padrão parece com isso:

```
gcloud container clusters create cluster1 --num-nodes=3 --region=us-central1
```

Há um grande número de parâmetros para o comando **gcloud container clusters create** que permitem especificar muitas configurações diferentes para um cluster. Para detalhes sobre os parâmetros, visite <https://cloud.google.com/sdk/gcloud/reference/container/clusters/create>.

O comando **gcloud container clusters create-auto** é usado para criar clusters GKE no modo piloto automático.

Implantando Pods de Aplicativos Agora que você criou um cluster, vamos implantar um aplicativo.

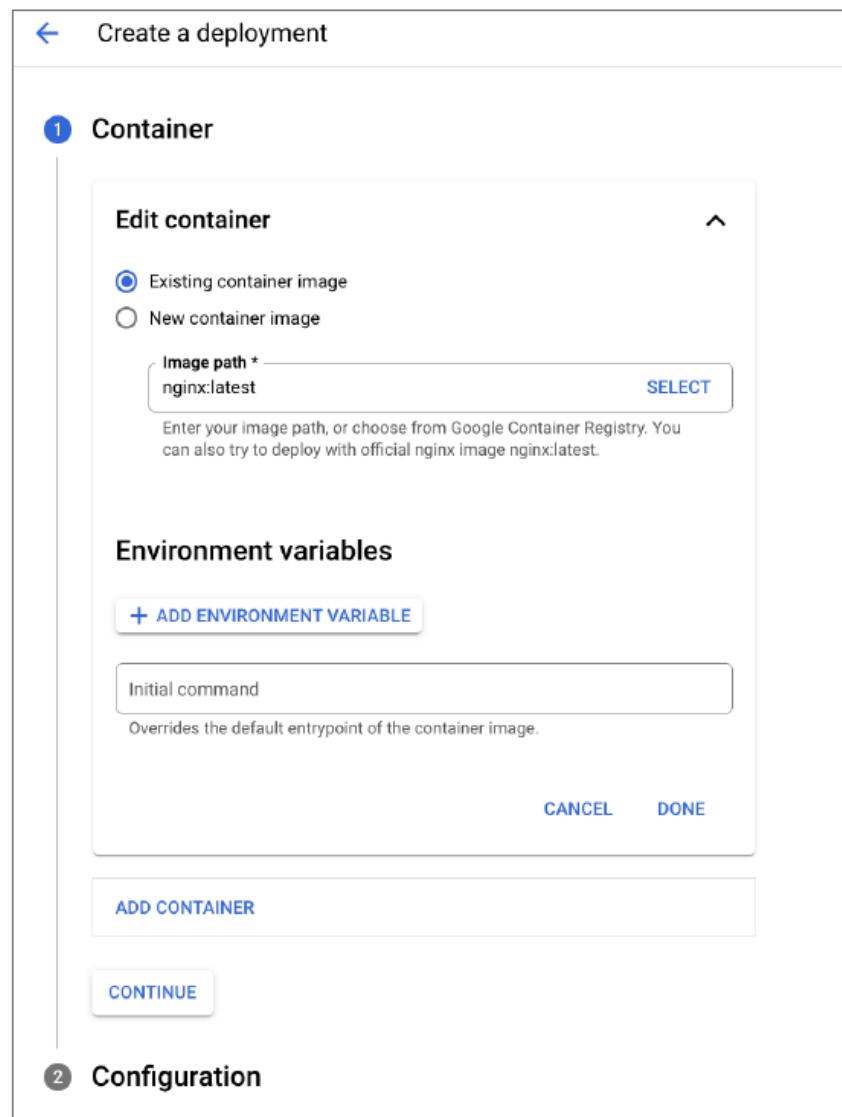
Na página Clusters do Kubernetes Engine no Cloud Console, selecione Criar Implantação. Um formulário como o mostrado na Figura 7.8 aparece. Use este formulário para especificar o seguinte:

- Imagem do contêiner
- Variáveis de ambiente
- Comando inicial

Após especificar os parâmetros iniciais, você pode continuar a adicionar parâmetros de configuração (veja a Figura 7.9):

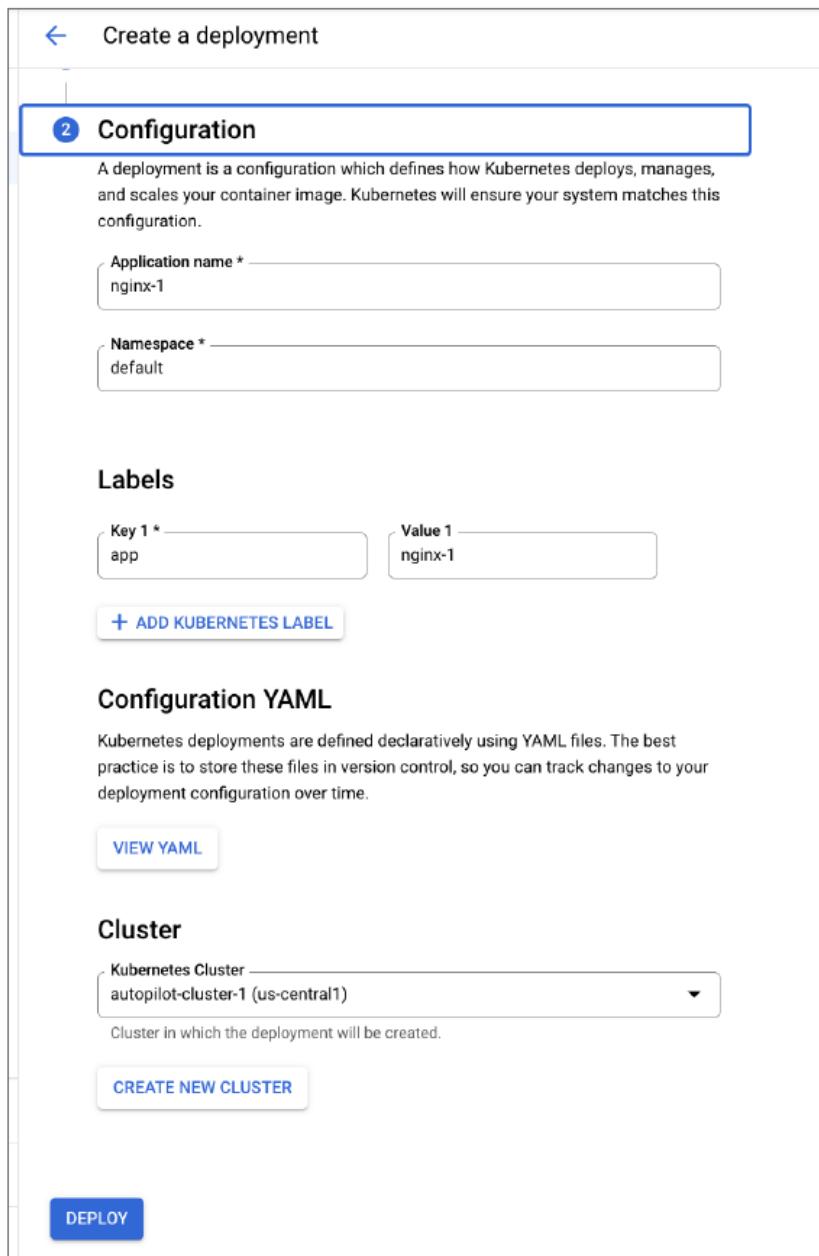
- Nome do aplicativo
- Namespace
- Etiquetas
- Cluster

FIGURE 7.8 The Create Deployment option provides a form to specify a container to run and an initial command to start the application running.



Depois de especificar uma implantação, você pode exibir a especificação YAML correspondente, que pode ser salva e usada para criar implantações a partir da linha de comando. Os elementos centrais do template do Kubernetes incluem apiVersion, kind, metadata e spec. A Listagem 7.1 mostra um exemplo de arquivo YAML de implantação. A saída é sempre exibida no formato YAML.

FIGURE 7.9 Configuring a deployment



Listing 7.1: Sample YAML configuration specification for a deployment

```
apiVersion: "apps/v1"
```

```
kind: "Deployment"
```

```
metadata:
```

```
name: "nginx-1"
```

```
namespace: "default"
```

```
labels:
```

```
app: "nginx-1"
```

```
spec:  
replicas: 3  
selector:  
matchLabels:  
app: "nginx-1"  
template:  
metadata:  
labels:  
app: "nginx-1"  
spec:  
containers:  
-name:  
"nginx-1"  
image: "nginx:latest"  
--  
-  
apiVersion:  
"autoscaling/v2beta1"  
kind: "HorizontalPodAutoscaler"  
metadata:  
name: "nginx-1-  
hpa-  
5fkn"  
namespace: "default"  
labels:  
app: "nginx-1"  
spec:  
scaleTargetRef:  
kind: "Deployment"  
name: "nginx-1"
```

```
apiVersion: "apps/v1"
minReplicas: 1
maxReplicas: 5
metrics:
  -type:
    "Resource"
  resource:
    name: "cpu"
targetAverageUtilization: 80
```

Além de instalar o Cloud SDK, você precisará instalar a ferramenta de linha de comando do Kubernetes, kubectl, para trabalhar com clusters a partir da linha de comando. Isso pode ser feito com o seguinte comando:

```
gcloud components install kubectl
```

Você pode então usar kubectl para executar uma imagem Docker em um cluster usando o comando kubectl run. Para executar um contêiner dentro de uma implantação, use o comando create deployment. Aqui está um exemplo:

```
kubectl create deployment app-deploy1 --image=app1 --port=8080
```

Isso executará uma imagem Docker chamada app1 e tornará sua rede acessível na porta 8080.

Se depois de algum tempo você quiser aumentar o número de réplicas na implantação, você pode usar o comando kubectl scale:

Bash

```
kubectl scale deployment app-deploy1 --replicas=5
```

Este exemplo criaria cinco réplicas.

Monitoramento do Kubernetes

O Cloud Operations Suite é o produto abrangente de monitoramento, registro em log e alertas do Google Cloud e inclui os serviços Cloud Monitoring e Cloud Logging, que podem ser usados para monitorar clusters Kubernetes.

O GKE fornece várias fontes de métricas de desempenho de aplicativos e sistemas, incluindo métricas do sistema, Managed Service for Prometheus e métricas de carga de trabalho. As métricas do sistema descrevem recursos de cluster de baixo nível como CPUs, memória e armazenamento. O Prometheus é um sistema de código aberto amplamente utilizado para coletar métricas de desempenho. O Managed Service for Prometheus é um serviço fornecido pelo Google Cloud para clientes que desejam usar o Prometheus, mas que não querem gerenciar a infraestrutura e aplicações que compõem o

Prometheus. Métricas de carga de trabalho são um conjunto de métricas depreciadas expostas pelas cargas de trabalho do GKE.

Ao criar um cluster, você pode indicar que as métricas sejam enviadas para o Cloud Monitoring e os logs sejam enviados para o Cloud Logging. Ambos são habilitados por padrão.

Resumo

Kubernetes Engine é um sistema de orquestração de contêineres para implantar aplicações para rodar em clusters. Kubernetes é estruturado com um único gerenciador de cluster e nós trabalhadores. Kubernetes utiliza o conceito de pods, ou instâncias executando um contêiner. É possível rodar múltiplos contêineres em um pod, mas isso geralmente é feito apenas quando os dois contêineres estão fortemente acoplados. ReplicaSets são controladores para garantir que o número correto de pods esteja rodando. Implantações são conjuntos de pods idênticos. StatefulSets são um tipo de implantação usado para aplicações com estado.

Clusters Kubernetes podem ser implantados através do Cloud Console ou usando comandos gcloud. Você implanta aplicações ao empacotar a aplicação em um contêiner e usando o console ou o comando kubectl para criar uma implantação que executa a aplicação no cluster.

O Cloud Operations Suite inclui o Cloud Monitoring e o Cloud Logging, que são usados para monitorar instâncias em clusters.

Exame

Entenda que o Kubernetes é um sistema de orquestração de contêineres. O Kubernetes Engine é um produto do Google Cloud que oferece Kubernetes aos clientes do Google Cloud. O Kubernetes gerencia contêineres que rodam em um conjunto de instâncias VM.

Entenda que o Kubernetes usa um plano de controle para gerenciar nós e cargas de trabalho. O Kubernetes usa o plano de controle para coordenar a execução e monitorar a saúde dos pods. Se houver um problema com um pod, o plano de controle pode corrigir o problema e reagendar o trabalho interrompido.

Saiba descrever pods. Pods são instâncias únicas de um processo em execução, serviços fornecem um nível de indireção entre pods e clientes chamando serviços nos pods, um ReplicaSet é um tipo de controlador que garante que o número correto de pods esteja rodando, e uma implantação é um conjunto de pods idênticos.

O Kubernetes pode ser implantado usando o Cloud Console ou usando comandos gcloud. Comandos gcloud manipulam o serviço Kubernetes Engine, enquanto comandos kubectl são usados para gerenciar o estado interno dos clusters a partir da linha de comando. O comando base para trabalhar com o Kubernetes Engine é gcloud container. Observe que gcloud e kubectl têm sintaxes de comando diferentes. Comandos kubectl especificam um verbo e depois um recurso, como em kubectl scale deployment ..., enquanto gcloud especifica um recurso antes do verbo, como em gcloud container clusters

create. Implantações são criadas usando o Cloud Console ou na linha de comando usando uma especificação YAML.

Saiba definir objetos do Kubernetes. Implantações são conjuntos de pods idênticos. StatefulSets são um tipo de implantação usada para aplicações com estado. O Kubernetes é monitorado usando o Cloud Operations. O Cloud Operations pode ser configurado para gerar alertas e notificar você em uma variedade de canais. Para monitorar o estado de um cluster, você pode criar uma política que monitora uma métrica, como a utilização da CPU, e ter notificações enviadas por e-mail ou outros canais.

Questões

1. Um novo engenheiro está pedindo esclarecimentos sobre quando é melhor usar Kubernetes e quando usar grupos de instâncias. Você aponta que Kubernetes utiliza grupos de instâncias. Qual é o propósito dos grupos de instâncias em um cluster Kubernetes?
 - A. Eles monitoram a saúde das instâncias.
 - B. Eles criam pods e implantações.
 - C. Eles criam conjuntos de VMs que podem ser gerenciados como uma unidade.
 - D. Eles criam alertas e canais de notificação.
2. Quais componentes são necessários em um cluster Kubernetes?
 - A. Um plano de controle e nós para executar cargas de trabalho.
 - B. Um plano de controle, nós para executar cargas de trabalho e nós de monitoramento para monitorar a saúde dos nós.
 - C. Nós do Kubernetes; todas as instâncias são iguais.
 - D. Instâncias com pelo menos seis vCPUs.
3. O que é um pod em Kubernetes?
 - A. Um conjunto de contêineres.
 - B. Código de aplicação implantado em um cluster Kubernetes.
 - C. Uma única instância de uma aplicação em execução em um cluster.
 - D. Um controlador que gerencia a comunicação entre clientes e serviços do Kubernetes.
4. Você desenvolveu uma aplicação que chama um serviço rodando em um cluster Kubernetes. O serviço roda em pods que podem ser terminados se estiverem insalubres e substituídos por outros pods que podem ter um endereço IP diferente. Como você deve codificar sua aplicação para garantir que ela funcione corretamente nessa situação?
 - A. Consulte o Kubernetes por uma lista de endereços IP de pods rodando o serviço que você usa.
 - B. Comunique-se com os Serviços do Kubernetes para que as aplicações não precisem ser acopladas a pods específicos.
 - C. Consulte o Kubernetes por uma lista de pods rodando o serviço que você usa.
 - D. Use um comando gcloud para obter os endereços IP necessários.
5. Você notou que o desempenho de uma aplicação degradou significativamente. Você fez recentemente algumas alterações de configuração nos recursos do seu cluster Kubernetes e suspeita que essas alterações alteraram o número de pods

rodando no cluster. Onde você procuraria detalhes sobre o número de pods que deveriam estar rodando?

A. Configuração de implantação.

B. Cloud Operations Suite.

C. Runtime do Contêiner.

D. Jobs.

6. Você está implantando uma aplicação de alta disponibilidade no Kubernetes Engine. Você quer manter a disponibilidade mesmo que haja uma grande interrupção de rede em um data center. Qual recurso do Kubernetes Engine você usaria?

A. Múltiplos grupos de instâncias

B. Cluster regional

C. Implantações regionais

D. Balanceamento de carga

7. Você quer escrever um script para implantar um cluster Kubernetes com GPUs. Você já implantou clusters antes, mas não tem certeza sobre todos os parâmetros necessários. Você precisa implantar esse script o mais rápido possível. Qual é uma maneira de desenvolver este script rapidamente?

A. Use o template de GPU no console da nuvem do Kubernetes Engine para gerar o comando gcloud para criar o cluster.

B. Pesquise na web por um script.

C. Reveja a documentação sobre parâmetros gcloud para adicionar GPUs.

D. Use um script existente e adicione parâmetros para anexar GPUs.

8. Qual comando gcloud cria um cluster chamado ch07-cluster-1 com quatro nós?

A. gcloud container clusters create ch07-cluster-1 --num-nodes=4

B. gcloud container clusters create ch07-cluster-1 --size=4

C. gcloud container clusters create ch07-cluster-1 --region-nodes=4

D. gcloud beta container clusters create ch07-cluster-1 4

9. Ao usar Criar Implantação do Console da Nuvem, qual dos seguintes não pode ser especificado para uma implantação?

A. Imagem do contêiner

B. Nome da aplicação

C. Tempo de Vida (TTL)

D. Comando inicial

10. Arquivos de configuração de implantação criados no Console da Nuvem usam qual tipo de formato de arquivo?

- A. CSV
- B. YAML
- C. TSV
- D. JSON

11. Qual comando é usado para rodar uma imagem Docker em um cluster?

- A. gcloud container run
- B. gcloud container clusters run
- C. kubectl run
- D. kubectl container run

12. Qual comando você usaria para ter 10 réplicas de uma implantação chamada ch07-app-deploy?

- A. kubectl upgrade deployment ch07-app-deploy --replicas=5
- B. gcloud containers deployment ch07-app-deploy --replicas=5
- C. kubectl scale deployment ch07-app-deploy --replicas=10
- D. kubectl scale deployment ch07-app-deploy --pods=5

13. Cloud Operations Suite é usado para quais operações em clusters Kubernetes?

- A. Apenas notificações
- B. Apenas monitoramento e notificações
- C. Apenas registro de logs
- D. Notificações, monitoramento e registro de logs

14. Você quer usar o Cloud Logging e o Cloud Monitoring com seus clusters GKE. O que você deve fazer para habilitar isso ao criar um cluster?

- A. Especifique os parâmetros --monitoring=True e --logging=True no comando gcloud container create cluster.
- B. Crie um grupo de nós e configure-o para monitoramento e registro de logs.
- C. Crie um namespace e configure-o para monitoramento e registro de logs.
- D. Nada; métricas e logs são enviados ao Cloud Logging e Cloud Monitoring por padrão.

15. Qual popular ferramenta de monitoramento de código aberto está disponível no Google Cloud como um serviço gerenciado?

- A. Prometheus
 - B. Apache Flink
 - C. MongoDB
 - D. Spark
16. Você quer criar um cluster no Kubernetes Engine e quer minimizar a quantidade de configuração e gestão de infraestrutura. Que tipo de cluster você criaria?
- A. Cluster em modo padrão
 - B. Cluster em modo Autopilot
 - C. Cluster em modo mínimo
 - D. Cluster em modo de template
17. Você quer o maior grau de controle sobre seu cluster Kubernetes. Que tipo de cluster você criaria?
- A. Cluster em modo padrão
 - B. Cluster em modo Autopilot
 - C. Cluster em modo mínimo
 - D. Cluster em modo de template
18. Você quer criar um cluster Kubernetes, mas não quer que o GKE atualize automaticamente o cluster. Como você configuraria o cluster?
- A. Com um canal de lançamento
 - B. Com um canal estático
 - C. Com vários grupos de nós
 - D. Com um ReplicaSet
19. Você está tentando executar comandos para iniciar uma implantação em um cluster Kubernetes. Os comandos não estão tendo efeito. Você suspeita que um componente do Kubernetes não está funcionando corretamente. Qual componente poderia ser o problema?
- A. A API do Kubernetes
 - B. Um StatefulSet
 - C. Comandos gcloud do Cloud SDK
 - D. ReplicaSet
20. Você implantou uma aplicação em um cluster Kubernetes. Você notou que vários pods estão carentes de recursos por um período de tempo e os pods são desligados. Quando os recursos estão disponíveis, novas instanciações desses pods são criadas. Os clientes ainda conseguem se conectar aos pods, mesmo que os novos

pods tenham endereços IP diferentes dos pods que foram terminados. Qual componente do Kubernetes torna isso possível?

- A. Serviços
- B. ReplicaSet
- C. Alertas
- D. StatefulSet

Capítulo 8

Gerenciando Clusters Kubernetes em Modo Padrão

ESTE CAPÍTULO COBRE O SEGUINTE OBJETIVO DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

✓✓ 4.2 Gerenciando recursos do Google Kubernetes Engine

Este capítulo descreve como realizar tarefas básicas de gerenciamento do Kubernetes, incluindo o seguinte:

- Visualizando o status dos clusters Kubernetes
- Visualizando repositórios de imagens e detalhes de imagens
- Adicionando, modificando e removendo nós
- Adicionando, modificando e removendo pods
- Adicionando, modificando e removendo serviços

Você verá como realizar cada uma dessas tarefas usando o Google Cloud Console e o Cloud SDK, que você pode usar localmente em suas máquinas de desenvolvimento, em máquinas virtuais do Google Cloud e usando o Cloud Shell.

Visualizando o Status de um Cluster Kubernetes

Assumindo que você criou um cluster usando os passos descritos no Capítulo 7, "Computando com Kubernetes", você pode visualizar o status de um cluster Kubernetes usando o Google Cloud Console ou os comandos gcloud.

Visualizando o Status dos Clusters Kubernetes Usando o Cloud Console

Começando da página inicial do Cloud Console, abra o menu de navegação clicando no ícone de três linhas empilhadas no canto superior esquerdo. Isso exibe a lista de serviços do Google Cloud, como mostrado na Figura 8.1.

Selecione o Kubernetes Engine da lista de serviços para abrir o submenu mostrado na Figura 8.2.

FIGURE 8.1 Navigation menu in Google Cloud Console

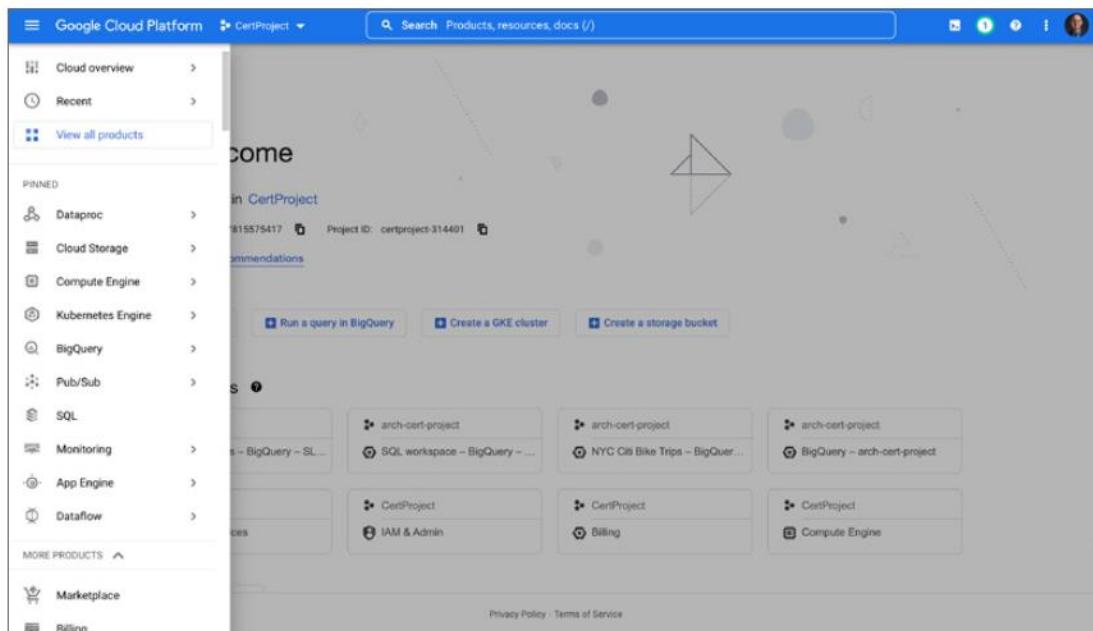
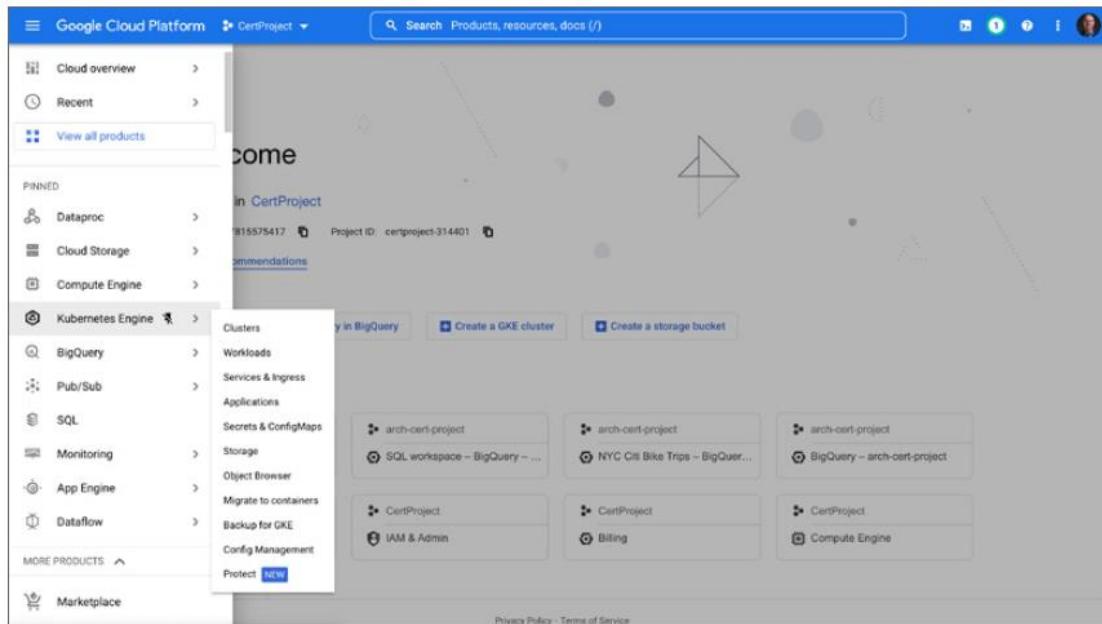


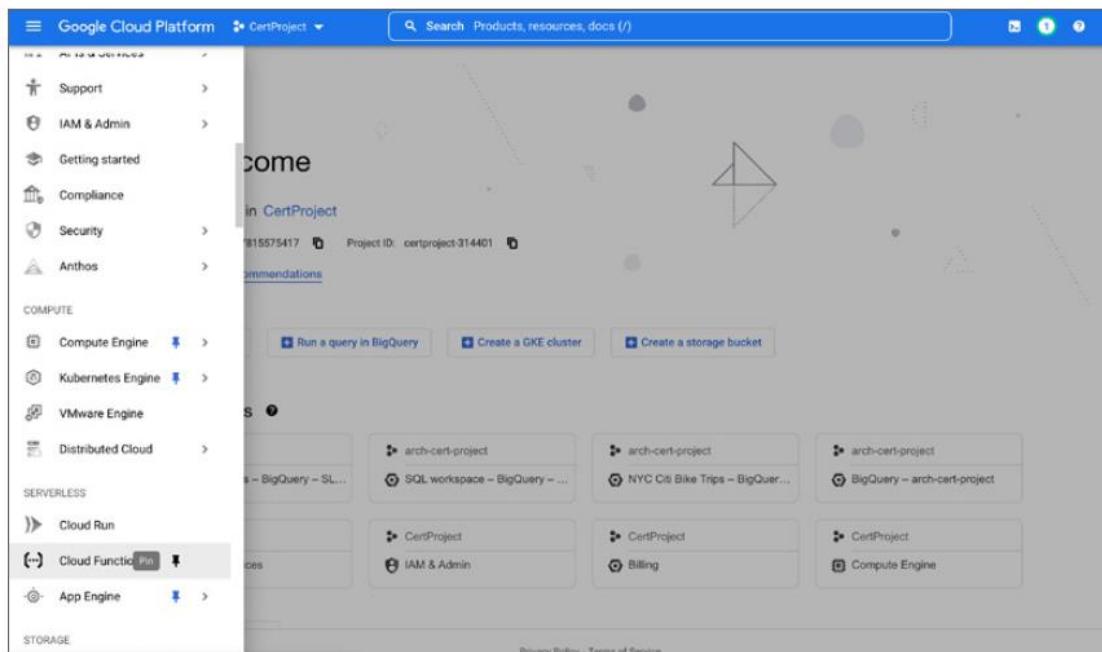
FIGURE 8.2 Selecting Kubernetes Engine from the navigation menu



Fixando Serviços no Topo do Menu de Navegação

Na Figura 8.2, o Kubernetes Engine foi fixado para que seja exibido no topo. Você pode fixar qualquer serviço no menu de navegação passando o mouse sobre o produto e clicando no ícone de alfinete que aparece, como mostrado na Figura 8.3. Nessa figura, o Compute Engine e o Kubernetes Engine já estão fixados, e o Cloud Functions pode ser fixado clicando no ícone de alfinete cinza.

FIGURE 8.3 Pinning a service to the top of the navigation menu



Após clicar no Kubernetes Engine no menu de navegação, você verá uma lista de clusters em execução, como mostrado na Figura 8.4, que mostra um único cluster chamado standard-cluster-1.

FIGURE 8.4 Example list of clusters in Kubernetes Engine

The screenshot shows the Kubernetes Engine interface. On the left, there's a sidebar with icons for Clusters, Workloads, Services & Ingress, Applications, Secrets & ConfigMaps, Storage, Object Browser, Migrate to containers, Backup for GKE, and Config Management. The main area has tabs for OVERVIEW, COST OPTIMIZATION, and a Filter input field. A table lists one cluster: standard-cluster-1, located in us-west1-a, with 3 nodes, 6 vCPUs, and 12 GB of memory. There are also columns for Notifications and Labels.

Status	Name	Location	Number of nodes	Total vCPUs	Total memory	Notifications	Labels
<input type="checkbox"/>	standard-cluster-1	us-west1-a	3	6	12 GB	-	-

Passe o mouse sobre o nome do cluster para destacá-lo, como na Figura 8.5, e clique no nome para exibir detalhes do cluster, como mostrado na Figura 8.6.

FIGURE 8.5 Click the name of a cluster to display its details.

This screenshot shows the same interface as Figure 8.4, but with the 'OVERVIEW' tab selected. The cluster 'standard-cluster-1' is highlighted with a blue selection bar. The table data remains the same as in Figure 8.4.

Status	Name	Location	Number of nodes
<input type="checkbox"/>	<u>standard-</u> <u>cluster-1</u>	us-west1-a	3

FIGURE 8.6 The first part of the cluster Details page describes the configuration of the cluster.

The screenshot shows the 'Clusters' section of the Kubernetes Engine interface. On the left is a sidebar with navigation links: Clusters (selected), Workloads, Services & Ingress, Applications, Secrets & ConfigMaps, Storage, Object Browser, Migrate to Containers, Backup for GKE (NEW), Security Posture, Config & Policy (Config selected), Marketplace, and Release Notes. The main area is titled 'standard-cluster-1'. It has tabs for DETAILS (selected), NODES, STORAGE, OBSERVABILITY, and LOGS. The DETAILS tab displays 'Cluster basics' with the following data:

	Value	Action
Name	standard-cluster-1	🔒
Location type	Zonal	🔒
Control plane zone	us-west1-a	🔒
Default node zones	us-west1-a	✍
Release channel	Regular channel	✍ UPGRADE AVAILABLE
Version	1.24.7-gke.900	
Total size	3	ⓘ
External endpoint	34.168.24.163 Show cluster certificate	✍
Internal endpoint	10.138.0.10 Show cluster certificate	🔒

Below this is a 'Automation' section with the following data:

	Value	Action
Maintenance window	Any time	✍
Maintenance exclusions	None	
Notifications	Disabled	✍
Vertical Pod Autoscaling	Disabled	✍

Clicar na opção Nós mostra detalhes dos grupos de nós e réplicas (veja a Figura 8.7). Na seção Grupos de Nós, você verá o número de nós, tipo de máquina, tipo de imagem e outros atributos. Na seção Nós, você verá os nós e seus status, CPUs solicitadas e alocáveis, memória e armazenamento.

FIGURE 8.7 Add-on and permission details for a cluster

The screenshot shows the Google Cloud Platform (GCP) Kubernetes Engine interface for a cluster named 'standard-cluster-1'. The 'NODES' tab is active. The 'Node Pools' section shows one pool named 'default-pool' with 3 nodes. The 'Nodes' section shows two nodes: 'gke-standard-cluster-1-default-pool-828d85db-1191' and 'gke-standard-cluster-1-default-pool-828d85db-1192', both in a 'Ready' state.

Name ↑	Status	Version	Number of nodes	Machine type	Image type	Autoscaling
default-pool	Ok	1.24.7-gke.900	3	e2-medium	Container-Optimized OS with containerd (cos_containerd)	Off

Name ↑	Status	CPU requested	CPU allocatable	Memory requested	Memory allocatable
gke-standard-cluster-1-default-pool-828d85db-1191	Ready	488 mCPU	940 mCPU	492.83 MB	2.95 GB
gke-standard-cluster-1-default-pool-828d85db-1192	Ready	528 mCPU	940 mCPU	534.29 MB	2.95 GB

A Figura 8.8 mostra detalhes exemplares dos grupos de nós, que são grupos de instâncias separadas executando em um cluster Kubernetes. Os detalhes nesta seção incluem a imagem de nó em execução nos nós, o tipo de máquina, o número total de vCPUs (listados como Núcleos Totais), o tipo de disco e se os nós são preempíveis.

Abaixo do nome do cluster há uma lista horizontal de várias opções: Detalhes, Nós, Armazenamento, Observabilidade e Logs. Até agora, descrevemos o conteúdo da página de Detalhes. Clique em Armazenamento para exibir informações como as mostradas na Figura 8.9, que exibe volumes persistentes e as classes de armazenamento usadas pelo cluster.

Este cluster não possui volumes persistentes, mas usa armazenamento padrão. Volumes persistentes são discos duráveis que são gerenciados pelo Kubernetes e implementados usando discos persistentes do Compute Engine. Uma classe de armazenamento é um tipo de armazenamento com um conjunto de políticas especificando qualidade de serviço, política de backup e um provisionador (que é um serviço que implementa o armazenamento).

FIGURE 8.8 Details about node pools in the cluster

default-pool	
Cluster	standard-cluster-1
Node version	1.21.11-gke.1100
Size	
Number of nodes	3
Autoscaling	Off
Node zones	us-west1-a
Nodes	
Image type	Container-Optimized OS with containerd (cos_containerd)
Machine type	e2-medium
Boot disk type	Standard persistent disk
Boot disk size (per node)	100 GB
Boot disk encryption	Google-managed
Provisioning Model	Standard
Networking	
Pod IP Address Range	10.36.0.0/14 (gke-standard-cluster-1-pods-d3f4a502) (inherited from standard-cluster-1)
Maximum Pods per Node	110 (inherited from standard-cluster-1)
Management	
Auto-upgrade	Enabled
Auto-repair	Enabled
Surge upgrade	Enabled
Max surge	1
Max unavailable	0

A seção de Observabilidade mostra métricas sobre o desempenho do cluster. Sob a opção Logs do menu de status do cluster, você pode ver um registro de mensagens, como mostrado na Figura 8.10. Clique no nome de um dos nós para ver informações detalhadas de status, como mostrado na Figura 8.11. Os detalhes do nó incluem utilização da CPU, consumo de memória e I/O de disco. Há também uma lista de pods rodando no nó.

Clique no nome de um pod para ver seus detalhes. A exibição do pod é semelhante à exibição do nó, com estatísticas de CPU, memória e disco. Detalhes de configuração incluem quando o pod foi criado, as etiquetas atribuídas, links para logs e o status (que é mostrado como Running na Figura 8.12).

Outros possíveis status são Pending, que indica que o pod está baixando imagens; Succeeded, que indica que o pod terminou com sucesso; Failed, que indica que pelo

menos um contêiner falhou; e Unknown, que significa que o plano de controle não pode alcançar o nó e o status não pode ser determinado.

Na parte inferior da exibição do pod há uma lista de contêineres em execução. Clique no nome de um contêiner para ver seus detalhes. A Figura 8.13 mostra os detalhes de um Pod. As informações incluem o status, o horário de início, o comando que está sendo executado e os volumes montados.

FIGURE 8.9 Storage information about a cluster

The screenshot shows the 'Storage' tab of the cluster details page. It includes sections for 'Storage classes' and 'Persistent volumes'.

Storage classes:

GKE automatically deploys and manages the Kubernetes Filestore Container Storage Interface (CSI) driver. Enable the CSI driver to add Filestore (NFS) storage. If enabled, Filestore storage classes will appear in the table below. [Learn more](#)

Name	Provisioner	Type	Zone
premium-rwo	pd.csi.storage.gke.io	pd-ssd	
standard	kubernetes.io/gce-pd	pd-standard	
standard-rwo	pd.csi.storage.gke.io	pd-balanced	

Persistent volumes:

No persistent volume to display. Use Cloud Shell for YAML file creation and kubectl operations.

FIGURE 8.10 Log of nodes in the cluster

The screenshot shows the 'Clusters' page in the Google Cloud Platform UI. A cluster named 'standard-cluster-1' is selected. The 'NODES' tab is active, showing a single node. The 'LOGS' tab is selected, displaying a list of log entries. The logs are categorized into 'CLUSTER LOGS' and 'AUTOSCALER LOGS'. The severity is set to 'Default'. The log entries are timestamped and show various system requests and responses from Kubernetes components like Apiservice Requests, kube-system, and kube-controller-manager.

Timestamp	Log Message
2022-05-27T15:04:11.884153Z	Kubernetes Apiservice Requests update kube-node-lease:gke-standard-cluster-1-de...
2022-05-27T15:04:12.262448Z	Kubernetes Apiservice Requests update kube-system:vpa-recommender system:vpa-recommender
2022-05-27T15:04:12.566576Z	Kubernetes Apiservice Requests update kube-system:cluster-metrics system:clustermetrics
2022-05-27T15:04:12.562854Z	Kubernetes Apiservice Requests update kube-system:ingress-gce-lock system:17-lb-controller
2022-05-27T15:04:12.634946Z	Kubernetes Apiservice Requests update kube-system:managed-certificate-control... system:managed-certificate-controller
2022-05-27T15:04:12.642112Z	Kubernetes Apiservice Requests update kube-system:managed-certificate-control... system:managed-certificate-controller
2022-05-27T15:04:12.717128Z	Kubernetes Apiservice Requests update kube-system:cluster-kubestore system:kubestore-collector
2022-05-27T15:04:12.966317Z	Kubernetes Apiservice Requests update kube-system:cluster-autoscaler system:cluster-autoscaler
2022-05-27T15:04:13.001985Z	Kubernetes Apiservice Requests update kube-system:snapshot-controller-leader system:snapshot-controller
2022-05-27T15:04:13.253194Z	Kubernetes Apiservice Requests update kube-system:kube-controller-manager system:kube-controller-manager
2022-05-27T15:04:13.258708Z	Kubernetes Apiservice Requests update kube-system:kube-scheduler system:kube-scheduler
2022-05-27T15:04:14.528855Z	Kubernetes Apiservice Requests update kube-system:clustermetrics system:clustermetrics
2022-05-27T15:04:14.577963Z	Kubernetes Apiservice Requests update kube-system:ingress-gce-lock system:17-lb-controller
2022-05-27T15:04:14.656655Z	Kubernetes Apiservice Requests update kube-system:managed-certificate-control... system:managed-certificate-controller
2022-05-27T15:04:14.736822Z	Kubernetes Apiservice Requests update kube-system:cluster-kubestore system:kubestore-collector
2022-05-27T15:04:15.269895Z	Kubernetes Apiservice Requests update kube-system:kube-controller-manager system:kube-controller-manager
2022-05-27T15:04:16.215054Z	Kubernetes Apiservice Requests update kube-system:vpa-recommender system:vpa-recommender
2022-05-27T15:04:16.591108Z	Kubernetes Apiservice Requests update kube-system:ingress-gce-lock system:17-lb-controller
2022-05-27T15:04:16.675467Z	Kubernetes Apiservice Requests update kube-system:managed-certificate-control... system:managed-certificate-controller

FIGURE 8.11 Example details of a node running in a Kubernetes cluster

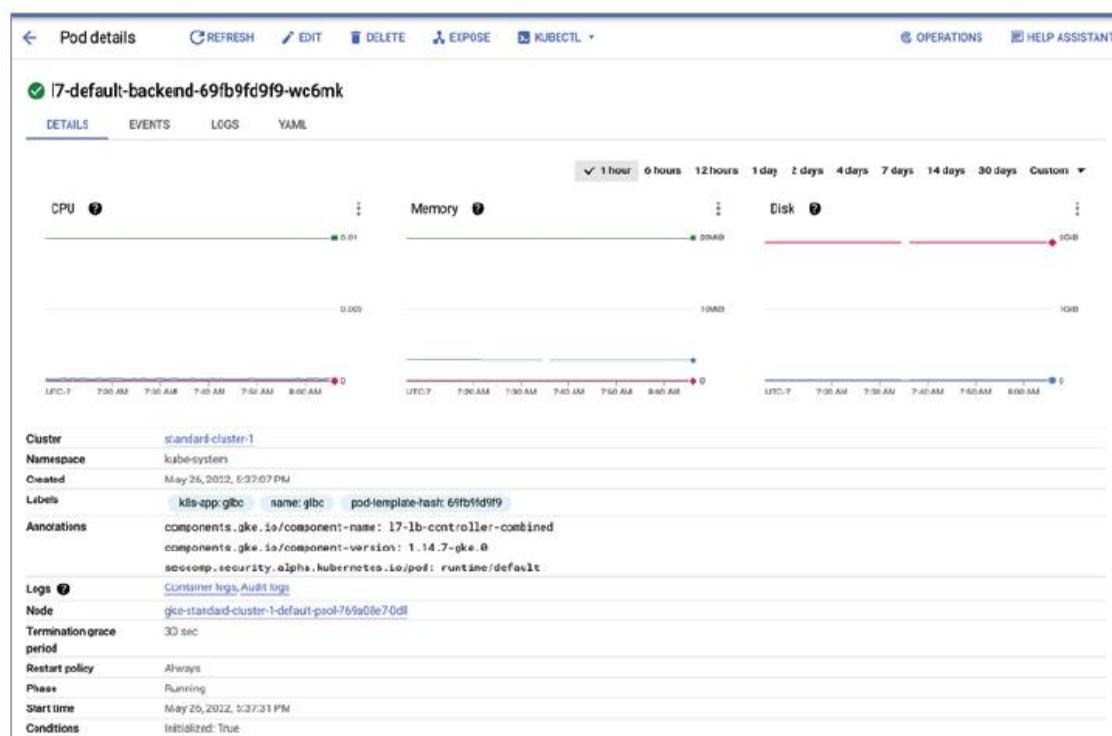
The screenshot shows the 'Node details' page for a node named 'gke-standard-cluster-1-default-pool-769a08e7-0dl1'. The 'SUMMARY' tab is active, displaying three line charts for CPU, Memory, and Disk usage over the last hour. The CPU chart shows usage at 0.5%, the Memory chart shows usage at 500MB, and the Disk chart shows usage at 0.0008GB. Below the charts, a timeline shows resource usage from UTC-7 to 8:00 AM. The 'PODS' tab is also visible, listing several pods running on the node, such as 'l7-default-backend-69fb9fdff9-wc6mk' and 'kube-dns-autoscaler-844c9d9448-888hr', all in a 'Running' state.

Name	Status	CPU requested	Memory requested	Storage requested	Namespace	Restarts	Created on
l7-default-backend-69fb9fdff9-wc6mk	Running	10 mCPU	20.97 MB	0 B	kube-system	0	May 26, 2022, 5:37:07 PM
kube-dns-autoscaler-844c9d9448-888hr	Running	20 mCPU	10.49 MB	0 B	kube-system	0	May 26, 2022, 5:37:16 PM
konnectivity-agent-autoscaler-6586f567c9-z72hd	Running	10 mCPU	10 MB	0 B	kube-system	0	May 26, 2022, 5:37:17 PM
konnectivity-agent-76d49f7d-h78n9	Running	10 mCPU	31.46 MB	0 B	kube-system	0	May 26, 2022, 5:37:17 PM
fluentbit-gke-7gh4x	Running	100 mCPU	209.72 MB	0 B	kube-system	0	May 26, 2022, 5:37:26 PM
pdcs-node-snslt	Running	10 mCPU	20.97 MB	0 B	kube-system	0	May 26, 2022, 5:37:27 PM

FIGURE 8.12 Pod status display, with the status Running

Name	Status	CPU requested	Memory requested	Storage requested	Namespace	Restarts	Created on
i7-default-backend-69fb9fd9f9-wc6mk	Running	10 mCPU	20.97 MB	0 B	kube-system	0	May 26, 2022, 5:37:07 PM
kube-dns-autoscaler-844c9d9448-88hr	Running	20 mCPU	10.49 MB	0 B	kube-system	0	May 26, 2022, 5:37:16 PM
konnectivity-agent-autoscaler-6b86f667c9-z72hd	Running	10 mCPU	10 MB	0 B	kube-system	0	May 26, 2022, 5:37:17 PM
konnectivity-agent-76d498f7d-h78n9	Running	10 mCPU	31.46 MB	0 B	kube-system	0	May 26, 2022, 5:37:17 PM
fluentbit-gke-7qh4x	Running	100 mCPU	209.72 MB	0 B	kube-system	0	May 26, 2022, 5:37:26 PM
pdcsi-node-srsit	Running	10 mCPU	20.97 MB	0 B	kube-system	0	May 26, 2022, 5:37:27 PM
gke-metrics-agent-xdc87	Running	3 mCPU	52.43 MB	0 B	kube-system	0	May 26, 2022, 5:37:27 PM
kube-dns-697dc8fc8b-2cwqg	Running	260 mCPU	115.34 MB	0 B	kube-system	0	May 26, 2022, 5:37:33 PM
kube-proxy-gke-standard-cluster-1-default-pool-769a08e7-0dl	Running	100 mCPU	0 B	0 B	kube-system	0	May 26, 2022, 5:38:15 PM

FIGURE 8.13 Details of a pod running on a node



Usando o Cloud Console, você pode listar todos os clusters e ver detalhes de sua configuração e status. Você pode então detalhar cada nó, pod e contêiner para ver seus detalhes.

Visualizando o Status dos Clusters Kubernetes Usando Cloud SDK e Cloud Shell

Você também pode usar a linha de comando para visualizar o status de um cluster. O comando `gcloud container clusters list` é usado para mostrar esses detalhes.

Para listar os nomes e informações básicas de todos os clusters, use este comando:

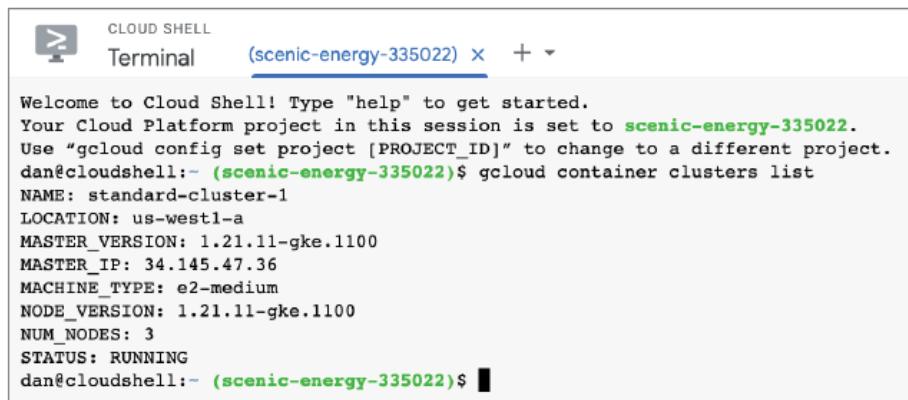
```
gcloud container clusters list
```

Isso produz a saída mostrada na Figura 8.14.

Por que os Comandos Não Começam com gcloud kubernetes?

Os comandos gcloud começam com a palavra gcloud seguida pelo nome do serviço, por exemplo, gcloud compute para comandos do Compute Engine e gcloud sql para comandos do Cloud SQL. Você poderia esperar que os comandos do Kubernetes Engine começassem com gcloud kubernetes, mas o serviço foi originalmente chamado de Google Container Engine. Em novembro de 2017, o Google renomeou o serviço para Kubernetes Engine, mas os comandos gcloud permaneceram os mesmos.

FIGURE 8.14 Example output from the `gcloud container clusters list` command



The screenshot shows a Cloud Shell terminal window with the title 'CLOUD SHELL' and 'Terminal'. The project is set to 'scenic-energy-335022'. The terminal output shows the following text:

```
Welcome to Cloud Shell! Type "help" to get started.  
Your Cloud Platform project in this session is set to scenic-energy-335022.  
Use "gcloud config set project [PROJECT_ID]" to change to a different project.  
dan@cloudshell:~ (scenic-energy-335022)$ gcloud container clusters list  
NAME: standard-cluster-1  
LOCATION: us-west1-a  
MASTER_VERSION: 1.21.11-gke.1100  
MASTER_IP: 34.145.47.36  
MACHINE_TYPE: e2-medium  
NODE_VERSION: 1.21.11-gke.1100  
NUM_NODES: 3  
STATUS: RUNNING  
dan@cloudshell:~ (scenic-energy-335022)$
```

Para visualizar os detalhes de um cluster, use o comando `gcloud container clusters describe`. Você precisará passar o nome de uma zona ou região usando o parâmetro `--zone` ou `--region`. Por exemplo, para descrever um cluster chamado `standard-cluster-1` localizado na zona `us-central1-a`, você usaria este comando:

```
gcloud container clusters describe --zone us-central1-a standard-cluster-1
```

Este comando exibirá detalhes como os mostrados na Figura 8.15 e Figura 8.16. Observe que o comando `describe` também exibe informações de autenticação, como certificado do cliente, nome de usuário e senha. Essas informações não são mostradas nas figuras.

FIGURE 8.15 Part 1 of the information displayed by the gcloud container clusters describe command

```
dan@cloudshell:~ (scenic-energy-335022)$ gcloud container clusters describe --zone us-west1-a standard-cluster-1
addonsConfig:
  dnsCacheConfig: {}
  gcePersistentDiskCsiDriverConfig:
    enabled: true
  horizontalPodAutoscaling: {}
  httpLoadBalancing: {}
  kubernetesDashboard:
    disabled: true
  networkPolicyConfig:
    disabled: true
  authenticatorGroupsConfig: {}
  autoscaling:
    autoscalingProfile: BALANCED
  binaryAuthorization: {}
  clusterIpv4Cidr: 10.36.0.0/14
  createTime: '2022-05-27T00:34:03+00:00'
  currentMasterVersion: 1.21.11-gke.1100
  currentNodeCount: 3
  currentNodeVersion: 1.21.11-gke.1100
  databaseEncryption:
    state: DECRYPTED
  defaultMaxPodsConstraint:
    maxPodsPerNode: '110'
  endpoint: 34.145.47.36
  id: d1f4a50283b04f22baf0a23b976d2869d7a1729135db43afbc13275b8a5d9ba9
  initialClusterVersion: 1.21.11-gke.1100
  instanceGroupOrls:
- https://www.googleapis.com/compute/v1/projects/scenic-energy-335022/zones/us-west1-a/instanceGroupManagers/gke-standard-cluster-1-default-pool-769a08e7-grp
ipAllocationPolicy:
  clusterIpv4Cidr: 10.36.0.0/14
  clusterIpv4CidrBlock: 10.36.0.0/14
  clusterSecondaryRangeName: gke-standard-cluster-1-pods-d1f4a502
  servicesIpv4Cidr: 10.40.0.0/20
  servicesIpv4CidrBlock: 10.40.0.0/20
  servicesSecondaryRangeName: gke-standard-cluster-1-services-d1f4a502
  useIAliases: true
  labelFingerprint: a9dc16a7
  legacyAbac: {}
location: us-west1-a
locations:
- us-west1-a
loggingConfig:
  componentConfig:
    enableComponents:
      - SYSTEM_COMPONENTS
      - WORKLOADS
  loggingService: logging.googleapis.com/kubernetes
```

FIGURE 8.16 Part 2 of the information displayed by the gcloud container clusters describe command

```
monitoringConfig:
  componentConfig:
    enableComponents:
      - SYSTEM_COMPONENTS
  monitoringService: monitoring.googleapis.com/kubernetes
name: standard-cluster-1
network: default
networkConfig:
  datapathProvider: LEGACY_DATAPATH
  defaultSnatStatus: {}
  network: projects/scenic-energy-335022/global/networks/default
  serviceExternalIpsConfig: {}
  subnetwork: projects/scenic-energy-335022/regions/us-west1/subnetworks/default
nodeConfig:
  diskSizeGb: 100
  diskType: pd-standard
  imageType: COS_CONTAINERD
  machineType: e2-medium
  metadata:
    disable-legacy-endpoints: 'true'
  oauthScopes:
    - https://www.googleapis.com/auth/devstorage.read_only
    - https://www.googleapis.com/auth/logging.write
    - https://www.googleapis.com/auth/monitoring
    - https://www.googleapis.com/auth/servicecontrol
    - https://www.googleapis.com/auth/service.management.readonly
    - https://www.googleapis.com/auth/trace.append
  serviceAccount: default
  shieldedInstanceConfig:
    enableIntegrityMonitoring: true
nodePoolAutoConfig: {}
nodePoolDefaults:
  nodeConfigDefaults: {}
nodePools:
- autoscaling: {}
  config:
    diskSizeGb: 100
    diskType: pd-standard
    imageType: COS_CONTAINERD
    machineType: e2-medium
    metadata:
      disable-legacy-endpoints: 'true'
    oauthScopes:
      - https://www.googleapis.com/auth/devstorage.read_only
      - https://www.googleapis.com/auth/logging.write
      - https://www.googleapis.com/auth/monitoring
      - https://www.googleapis.com/auth/servicecontrol
      - https://www.googleapis.com/auth/service.management.readonly
```

Para listar informações sobre nós e pods, use o comando kubectl. Primeiro, você precisa garantir que tenha um arquivo kubeconfig configurado corretamente, que contém informações sobre como se comunicar com a API do cluster. Execute o comando gcloud container clusters get-credentials com o nome de uma zona ou região e o nome de um cluster. Aqui está um exemplo:

```
gcloud container clusters get-credentials --zone us-central1-a standard-cluster-1
```

Este comando configurará o arquivo kubeconfig em um cluster chamado standard-cluster-1 na zona us-central1-a. A Figura 8.17 mostra um exemplo de saída desse comando, que inclui o status da busca e configuração dos dados de autenticação.

FIGURE 8.17 Example output of the get-credentials command

```
dan@cloudshell:~ (scenic-energy-335022)$ gcloud container clusters get-credentials --zone us-west1-a standard-cluster-1
Fetching cluster endpoint and auth data.
kubeconfig entry generated for standard-cluster-1.
dan@cloudshell:~ (scenic-energy-335022)$
```

Você pode listar os nós em um cluster usando o seguinte:

kubectl get nodes

Este comando produz uma saída como a mostrada na Figura 8.18, que mostra o status de três nós.

Da mesma forma, para listar pods, use o seguinte comando:

kubectl get pods

FIGURE 8.18 Example output of the kubectl get nodes command

```
CLOUD SHELL
Terminal (scenic-energy-335022) x + ~

dan@cloudshell:~ (scenic-energy-335022)$ kubectl get nodes
W0527 15:18:44.254844    1253 gcp.go:120] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.25+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME           STATUS   ROLES   AGE     VERSION
gke-standard-cluster-1-default-pool-769a08e7-0d11 Ready    <none>  14h    v1.21.11-gke.1100
gke-standard-cluster-1-default-pool-769a08e7-bz3h Ready    <none>  14h    v1.21.11-gke.1100
gke-standard-cluster-1-default-pool-769a08e7-v63z Ready    <none>  14h    v1.21.11-gke.1100
dan@cloudshell:~ (scenic-energy-335022)$
```

Este comando produz uma saída como a mostrada na Figura 8.19, que lista os pods e seus status. Para mais detalhes sobre nós e pods, use estes comandos:

kubectl describe nodes

kubectl describe pods

FIGURE 8.19 Example output of the kubectl get pods command

```
CLOUD SHELL
Terminal (scenic-energy-335022) x + ~

dan@cloudshell:~ (scenic-energy-335022)$ kubectl get pods -n kube-system
W0527 15:40:08.001134    1454 gcp.go:120] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.25+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME          READY   STATUS    RESTARTS   AGE
event-exporter-gke-5479fd58c8-ng2cd   2/2     Running   0          15h
fluentbit-gke-7hd6g      2/2     Running   0          15h
fluentbit-gke-7gh4x      2/2     Running   0          15h
fluentbit-gke-hlgnp      2/2     Running   0          15h
gke-metrics-agent-8g5zv   1/1     Running   0          15h
gke-metrics-agent-bwsj2   1/1     Running   0          15h
gke-metrics-agent-xdc87   1/1     Running   0          15h
konnectivity-agent-76d498f7d-db1qc   1/1     Running   0          15h
konnectivity-agent-76d498f7d-h78n9   1/1     Running   0          15h
konnectivity-agent-76d498f7d-tq9w5   1/1     Running   0          15h
konnectivity-agent-autoscaler-6b86f667c9-z72hd 1/1     Running   0          15h
kube-dns-697dc8fcfb-2cwqg      4/4     Running   0          15h
kube-dns-697dc8fcfb-46vcc      4/4     Running   0          15h
kube-dns-autoscaler-844cd9448-888hr  1/1     Running   0          15h
kube-proxy-gke-standard-cluster-1-default-pool-769a08e7-0d11 1/1     Running   0          15h
kube-proxy-gke-standard-cluster-1-default-pool-769a08e7-bz3h  1/1     Running   0          15h
kube-proxy-gke-standard-cluster-1-default-pool-769a08e7-v63z  1/1     Running   0          15h
17-default-backend-69fb9fd9f9-wc6mk   1/1     Running   0          15h
metrics-server-v0.4.5-bbb794dcc-8fk6g   2/2     Running   0          15h
pdcsi-node-pzsqs      2/2     Running   0          15h
pdcsi-node-r4w5s       2/2     Running   0          15h
pdcsi-node-sxslt      2/2     Running   0          15h
dan@cloudshell:~ (scenic-energy-335022)$
```

As Figuras 8.20 e 8.21 mostram listagens parciais dos resultados. Observe que o comando kubectl describe pods também inclui informações sobre contêineres, nomes, etiquetas, condições, endereços de rede e informações do sistema.

FIGURE 8.20 Partial listing of the details shown by the kubectl describe nodes command

Name:	gke-standard-cluster-1-default-pool-769a08e7-0d11				
Roles:	<none>				
Labels:	beta.kubernetes.io/arch=amd64 beta.kubernetes.io/instance-type=e2-medium beta.kubernetes.io/os=linux cloud.google.com/gke-boot-disk=pd-standard cloud.google.com/gke-container-runtime=containerd cloud.google.com/gke-nodepool=default-pool cloud.google.com/gke-os-distribution=cos cloud.google.com/machine-family=e2 failure-domain.beta.kubernetes.io/region=us-west1 failure-domain.beta.kubernetes.io/zone=us-west1-a kubernetes.io/arch=amd64 kubernetes.io/hostname=gke-standard-cluster-1-default-pool-769a08e7-0d11 kubernetes.io/os=linux node.kubernetes.io/instance-type=e2-medium topology.gke.io/zone=us-west1-a topology.kubernetes.io/region=us-west1 topology.kubernetes.io/zone=us-west1-a				
Annotations:	controller.googleapis.com/instance_id: 3507931133704489540 controller.googleapis.com/instance_id: 3507931133704489540 (pd.csi.storage.gke.io)" "projects/scenic-energy-33502/zones/us-west1-a/instances/gke-standard-cluster-1-default-pool-769a08e7-0d11" node.alpha.kubernetes.io/ttl: 0 node.gke.io/last-applied-node-labels: cloud.google.com/gke-boot-disk=pd-standard,cloud.google.com/gke-container-runtime=containerd,cloud.google.com/gke-nodepool=default-pool,cl... node.gke.io/last-applied-node-labels: volumes.kubernetes.io/controller-managed-attach-detach: true				
CreationTimestamp:	Fri, 27 May 2022 08:37:28 +0000				
Taints:	<none>				
Unschedulable:	false				
Leases:					
HeldByControllerIdentity:	gke-standard-cluster-1-default-pool-769a08e7-0d11				
AcquireTime:	<unset>				
ReleaseTime:	Fri, 27 May 2022 15:41:29 +0000				
Conditions:					
Type	Status	LastHeartbeatTime	LastTransitionTime	Reason	Message
---	---	---	---	---	---
CorruptDockerOverlay2	False	Fri, 27 May 2022 15:39:04 +0000	Fri, 27 May 2022 09:37:27 +0000	NoCorruptDockerOverlay2	docker overlay2 is functioning properly
FrequentUnregisterNetDevice	False	Fri, 27 May 2022 15:39:04 +0000	Fri, 27 May 2022 09:37:27 +0000	NoFrequentUnregisterNetDevice	node is functioning properly
FrequentDockerRestart	False	Fri, 27 May 2022 15:39:04 +0000	Fri, 27 May 2022 09:37:27 +0000	NoFrequentDockerRestart	docker is functioning properly
FrequentContainerRestart	False	Fri, 27 May 2022 15:39:04 +0000	Fri, 27 May 2022 09:37:27 +0000	NoFrequentContainerRestart	containerd is functioning properly
KernelDeadlock	False	Fri, 27 May 2022 15:39:04 +0000	Fri, 27 May 2022 09:37:27 +0000	KernelHasNoDeadlock	Kernel has no deadlock
ReadonlyFilesystem	False	Fri, 27 May 2022 15:39:04 +0000	Fri, 27 May 2022 09:37:27 +0000	FilesystemIsNotReadOnly	Filesystem is not read-only
NetworkUnavailable	False	Fri, 27 May 2022 09:37:26 +0000	Fri, 27 May 2022 09:37:26 +0000	RouteCreated	NodeController create implicit route
MemoryPressure	False	Fri, 27 May 2022 15:37:25 +0000	Fri, 27 May 2022 09:37:10 +0000	KubeletHasSufficientMemory	Kubelet has sufficient memory available
DiskPressure	False	Fri, 27 May 2022 15:37:25 +0000	Fri, 27 May 2022 09:37:10 +0000	KubeletHasNoDiskPressure	Kubelet has no disk pressure

FIGURE 8.21 Partial listing of the details shown by the kubectl describe pods command

Name:	event-exporter-gke-5479fd58c8-ng2cd
Namespace:	kube-system
Priority:	0
Node:	gke-standard-cluster-1-default-pool-769a08e7-br3h/10.138.0.2
Start Time:	Fri, 27 May 2022 00:37:35 +0000
Labels:	k8s-app=event-exporter pod-template-hash=5479fd58c8 version=v0.3.5
Annotations:	components.gke.io/component-name: event-exporter components.gke.io/component-version: 1.0.10
Status:	Running
IP:	10.36.0.3
IPs:	IP: 10.36.0.3
Controlled By:	ReplicaSet/event-exporter-gke-5479fd58c8
Containers:	
event-exporter:	
Container ID:	/19fb483de3fcecdb8dbda3b89434599940c8f0dca35224e42f92da2539b85fe8
Image:	gke.gcr.io/event-exporter:v0.3.5-gke.0
Image ID:	gke.gcr.io/event-exporter@sha256:09e908d7ea0020f47cc1279caef6ce1bd0cddeb9c9493b8550aace1d3c82c7c
Port:	<none>
Host Port:	<none>
Command:	/event-exporter --sink-opta--stackdriver-resource-model=new --endpoint=https://logging.googleapis.com
State:	Running
Started:	Fri, 27 May 2022 00:37:49 +0000
Ready:	True
Restart Count:	0
Environment:	<none>
Mounts:	/var/run/secrets/kubernetes.io/serviceaccount from kube-api-access-lfvlt (ro)
Prometheus-to-sd-exporter:	
Container ID:	/e6a0d73252606f2ff09eb0dc5169418d61f7969a0c9cb3ede709913b736ffcl
Image:	gke.gcr.io/prometheus-to-sd:v0.10.0-gke.0
Image ID:	gke.gcr.io/prometheus-to-sd@sha256:c5e12480a4319905e39ed249dc43a7672e99f7ef94a9928be40cf2f418f62f
Port:	<none>
Host Port:	<none>
Command:	/monitor --stackdriver-prefix=container.googleapis.com/internal/addons --api-overrides=https://monitoring.googleapis.com/ --source=event_exporter:http://localhost:8080?whitelisted=stackdriver_sink_received_entry_count,stackdriver_sink_request_count,stackdriver_sink_successfully_sent_entry_count --pod-id=\$POD_NAME --namespace-id=\$POD_NAMESPACE --node-name=\$NODE_NAME
State:	Running

Para visualizar o status dos clusters a partir da linha de comando, use os comandos gcloud container, mas para obter informações sobre objetos gerenciados pelo Kubernetes, como nós, pods e contêineres, use o comando kubectl.

Adicionando, Modificando e Removendo Nós

Você pode adicionar, modificar e remover nós de um cluster usando o Cloud Console ou o Cloud SDK em seu ambiente local, em uma máquina virtual do Google Cloud ou no Cloud Shell.

Adicionando, Modificando e Removendo Nós com o Cloud Console

No Cloud Console, navegue até a página do Kubernetes Engine e exiba uma lista de clusters. Clique no nome de um cluster para exibir seus detalhes, como mostrado na Figura 8.22.

FIGURE 8.22 Details of a cluster in Cloud Console

The screenshot shows the 'Clusters' page in the Google Cloud Console. At the top, there's a navigation bar with 'Clusters', 'EDIT', 'DELETE', 'ADD NODE POOL', 'OPERATIONS', and 'HELP ASSISTANT'. Below the navigation, the cluster name 'standard-cluster-1' is displayed with a checkmark icon. The 'DETAILS' tab is selected, showing the following sections:

- Cluster basics**:

Name	standard-cluster-1	🔒
Location type	Zonal	🔒
Control plane zone	us-central1-c	🔒
Default node zones <small>?</small>	us-central1-c	✍
Release channel	Regular channel	✍ UPGRADE AVAILABLE
Version	1.22.8-gke.201	
Total size	3	ⓘ
Endpoint	34.133.226.46	🔒
Show cluster certificate		
- Automation**:

Maintenance window	Any time	✍
Maintenance exclusions	None	✍
Notifications	Disabled	✍
Vertical Pod Autoscaling	Disabled	✍
Node auto-provisioning	Disabled	✍
Autoscaling profile	Balanced	✍
- Networking**:

Private cluster	Disabled	🔒
Network	default	🔒
Subnet	default	🔒
VPC-native traffic routing	Enabled	🔒
Cluster pod address range (default)	10.124.0.0/14	🔒 ▾
Maximum pods per node	110	🔒
Service address range	10.0.0.0/20	🔒
Intranode visibility	Disabled	✍
NodeLocal DNSCache	Disabled	✍
HTTP Load Balancing	Enabled	✍

Selecione a aba Nós para exibir as seções Grupos de Nós e Nós. A seção Grupos de Nós lista o nome, status, versão do GKE, número de nós, tipo de máquina e tipo de imagem. Também indica se o Escalonamento Automático está habilitado no grupo de nós. Selecione a opção Editar para alterar o número de nós no grupo de nós. A Figura 8.23 mostra detalhes de um grupo de nós.

FIGURE 8.23 Details of a node pool in Cloud Console

Name	Status	Version	Number of nodes	Machine type	Image type	Autoscaling
default-pool	Ok	1.22.8-gke.201	3	e2-medium	Container-Optimized OS with containerd (cos_containerd)	Off

Para adicionar nós, aumente o tamanho para o número de nós que você gostaria. Para remover nós, diminua o tamanho para o número de nós que você gostaria de ter.

Adicionando, Modificando e Removendo Nós com o Cloud SDK e Cloud Shell

O comando para adicionar ou modificar nós é `gcloud container clusters resize`. O comando leva três parâmetros:

- nome do cluster
- nome do grupo de nós
- tamanho do cluster

Por exemplo, suponha que você tenha um cluster chamado `standard-cluster-1` rodando um grupo de nós chamado `default-pool`. Para aumentar o tamanho do grupo de nós de 3 para 5, use este comando:

```
gcloud container clusters resize standard-cluster-1 --node-pool default-pool --num-nodes 5 --region=us-central1
```

O número de nós que você especificar no comando será o número de nós no grupo se você estiver usando um cluster zonal. Se você estiver usando um cluster regional, o número de nós será o número de nós para cada zona em que o grupo de nós está.

Uma vez que um cluster foi criado, você pode modificá-lo usando o comando `gcloud container clusters update`. Por exemplo, para habilitar o Escalonamento Automático, use o comando de atualização para especificar o número máximo e mínimo de nós. O comando para atualizar um cluster chamado `standard-cluster-1` rodando em um grupo de nós chamado `default-pool` é o seguinte:

```
gcloud container clusters update standard-cluster-1 --enable-autoscaling --min-nodes 1 --max-nodes 5 --zone us-central1-a --node-pool default-pool
```

Mantendo-se de Acordo com a Demanda com o Escalonamento Automático

Muitas vezes é difícil prever a demanda em um serviço. Mesmo que haja padrões regulares, como grandes trabalhos em lote executados durante horários fora do expediente, pode haver variação em quando esses picos de carga ocorrem. Em vez de continuar manualmente alterando o número de vCPUs em um cluster, habilite o Escalonamento Automático para adicionar ou remover nós automaticamente conforme necessário, com base na demanda. O Escalonamento Automático pode ser habilitado ao criar clusters com o Cloud Console ou gcloud. Esta abordagem é mais resiliente a picos inesperados e mudanças nos padrões de uso de pico a longo prazo. Isso também ajudará a otimizar o custo do seu cluster, não executando muitos servidores quando não necessário. Também ajudará a manter o desempenho, tendo nós suficientes para atender à demanda.

FIGURE 8.24 Deployment list of a cluster

The screenshot shows the 'Workloads' section of the Google Cloud Platform. At the top, there are buttons for REFRESH, DEPLOY, and DELETE, along with links for OPERATIONS and HELP ASSISTANT. Below these are dropdown menus for 'Cluster' and 'Namespace', and buttons for RESET and SAVE. A descriptive text box states: 'Workloads are deployable units of computing that can be created and managed in a cluster.' Below this, there are two tabs: 'OVERVIEW' (which is selected) and 'COST OPTIMIZATION'. Under 'OVERVIEW', there is a filter bar with 'Is system object : False' and a 'Filter workloads' input field. The main table lists one deployment:

	Name	Status	Type	Pods	Namespace	Cluster
<input type="checkbox"/>	nginx-1	OK	Deployment	3/3	default	standard-cluster-1

Clique no nome do deployment que você deseja modificar; um formulário é exibido com detalhes (veja Figura 8.25). Clique no nome de um pod na seção Managed Pods (veja Figura 8.26) para exibir detalhes do pod. Note que há um botão que permite deletar o pod na barra de menu horizontal no topo da página. Novamente, isso não é uma melhor prática em geral e os pods devem ser gerenciados pelo Kubernetes.

Selecione a opção Ações do ícone de três pontos verticais para exibir Ações, depois selecione Escalar para exibir uma caixa de diálogo que permite definir um novo tamanho para o workload, conforme mostrado na Figura 8.27. Neste exemplo, o número de réplicas foi alterado para 2.

Você também pode ter o Kubernetes adicionando e removendo réplicas (e pods) automaticamente dependendo da necessidade especificando o Autoscaling. Você pode escolher Autoscale no menu Ações, que é mostrado na Figura 8.28. No formulário resultante, você pode especificar um número mínimo e máximo de réplicas para executar.

O menu Ações também fornece opções para expor um serviço em uma porta, conforme mostrado na Figura 8.29, e especificar parâmetros para controlar atualizações contínuas do código implantado, conforme mostrado na Figura 8.30. Os parâmetros incluem os segundos mínimos de espera antes de considerar o pod atualizado, o número

máximo de pods acima do tamanho alvo permitido e o número máximo de pods indisponíveis.

FIGURE 8.25 Multiple forms contain details of a deployment and include a menu of actions you can perform on the deployment.

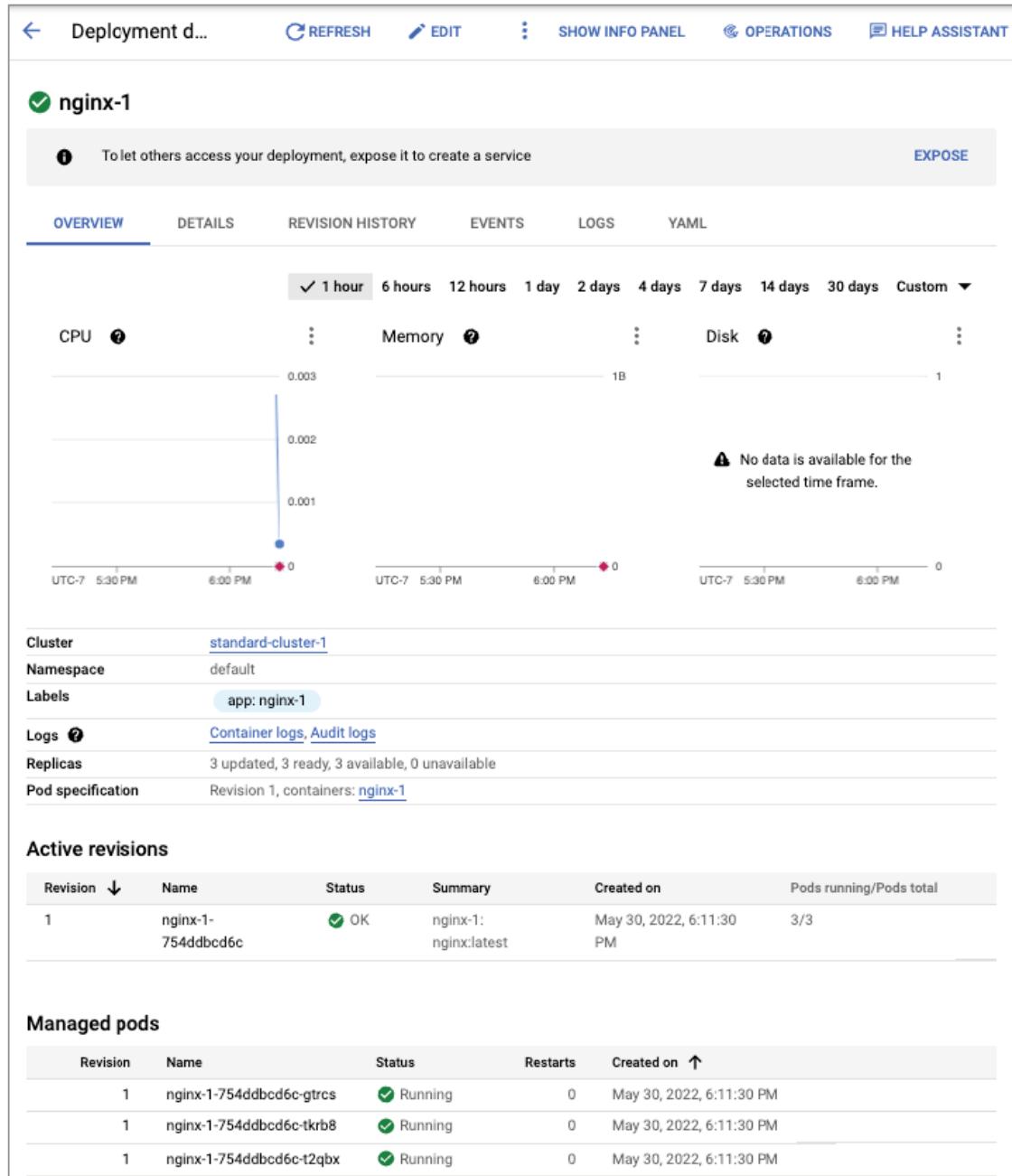


FIGURE 8.26 Details of a pod running in GKE

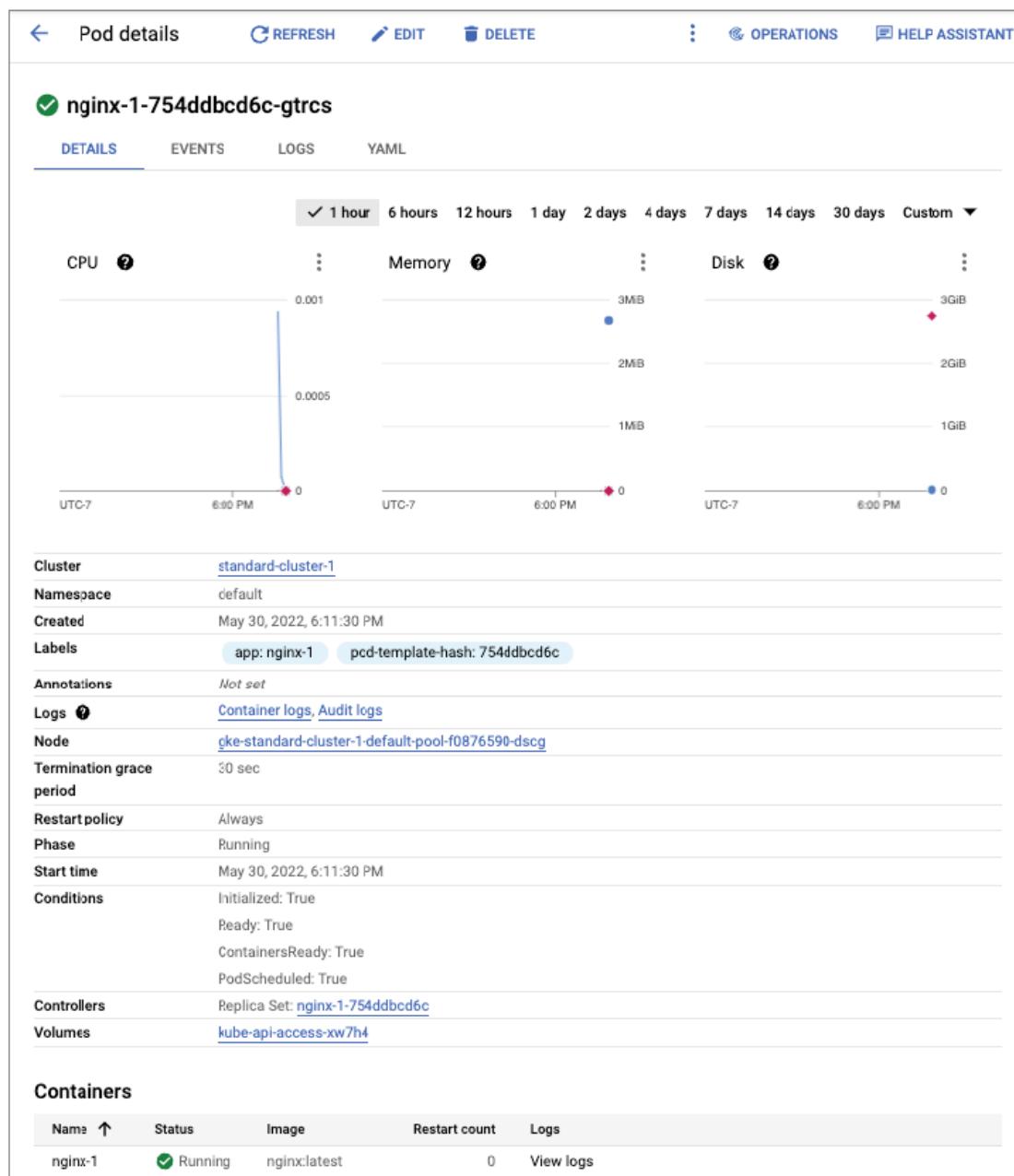


FIGURE 8.27 Set the number of replicas for a deployment.

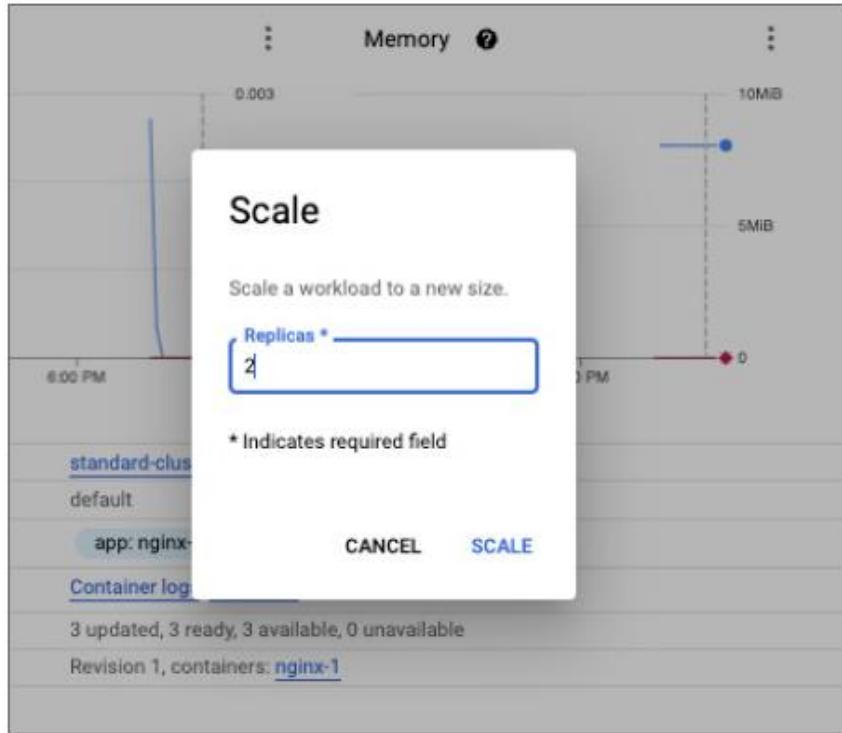
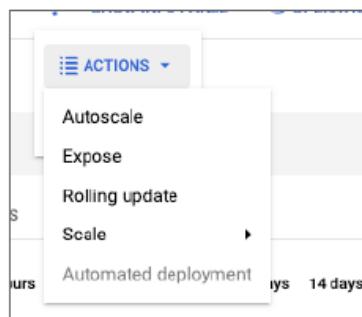


FIGURE 8.28 Enable Autoscaling to automatically add and remove replicas as needed depending on load.



Adicionando, Modificando e Removendo Pods com Cloud SDK e Cloud Shell

Trabalhar com pods no Cloud SDK e Cloud Shell é feito trabalhando com deployments; deployments foram explicados anteriormente na seção "Adicionando, Modificando e Removendo Pods com Cloud Console". Você pode usar o comando `kubectl` para trabalhar com deployments.

FIGURE 8.29 Form to expose services running on pods

The screenshot shows the Kubernetes UI for a deployment named "nginx-1". The "OVERVIEW" tab is selected. A modal window titled "Expose" is open, prompting the user to "Expose a resource's Pods using a Kubernetes Service". Inside the modal, there is a "Port mapping" section with fields for "Port 1" (set to 80), "Target port 1" (set to TCP), and "Protocol 1" (set to TCP). Below this is a dropdown for "Service type" set to "Cluster IP". A note at the bottom indicates that the asterisk (*) marks required fields. At the bottom right of the modal are "CANCEL" and "EXPOSE" buttons. The background shows resource details like CPU usage, disk space (10GiB and 5GiB), and a timeline. A table at the bottom lists active revisions, showing one revision named "nginx-1-754ddbcdec" with status "OK" and summary "nginx-1: nginx: datest".

Revision	Name	Status	Summary	Created on	Pods running/Pods total
1	nginx-1-754ddbcdec	OK	nginx-1: nginx: datest	May 30, 2022, 6:11:30 PM	3/3

Para listar deployments, use o seguinte comando:

```
kubectl get deployments
```

Para adicionar e remover pods, altere a configuração de deployments usando o comando kubectl scale deployment. Para este comando, você tem que especificar o nome do deployment e o número de réplicas. Por exemplo, para definir o número de réplicas para 5 para um deployment chamado nginx-1, use isto:

```
kubectl scale deployment nginx-1 --replicas=5
```

FIGURE 8.30 Form to specify parameters for rolling updates of code running in pods

The screenshot shows the Kubernetes UI for managing a deployment named 'nginx-1'. On the left, there's a sidebar with options like Cluster, Namespace, Labels, Logs, Replicas, and Pod specifications. The main area has tabs for OVERVIEW, DETAILS, REVISION HISTORY, EVENTS, LOGS, and YAML. The OVERVIEW tab is selected. A large central box contains a 'Rolling update' configuration dialog. This dialog includes fields for 'Minimum seconds ready' (0), 'Maximum surge' (25%), and 'Maximum unavailable' (25%). It also shows resource limits: 'CPU' with '1 hour' and 'Disk' with '6.00'. Below this is a 'Container images' section with the image 'Image of nginx-1 * nginx:latest'. At the bottom of the dialog are 'CANCEL' and 'UPDATE' buttons. The 'REVISION HISTORY' section shows one revision (Revision 1) with status 'OK'. The 'Managed pods' section lists pods like 'nginx-1-754ddbed6c'. The entire interface is in UTC-7.

Para ter o Kubernetes gerenciando o número de pods baseado na carga, use o comando `autoscale`. O comando seguinte adicionará ou removerá pods conforme necessário para atender à demanda baseada na utilização da CPU. Se o uso da CPU exceder 80%, até 10 pods ou réplicas adicionais serão adicionados. O deployment sempre terá pelo menos um pod ou réplica.

```
kubectl autoscale deployment nginx-1 --max=10 --min=1 --cpu-percent=80
```

Para remover um deployment, use o comando `delete deployment` assim:

```
kubectl delete deployment nginx-1
```

Serviços versus serviços

Na próxima seção, discutimos uma abstração do Kubernetes conhecida como Services. Um Serviço do Kubernetes é uma maneira de expor uma aplicação em um conjunto de pods para outras aplicações e usuários em uma rede. O termo serviços

também é usado em um sentido mais genérico como sinônimo de uma aplicação. Por exemplo, uma aplicação que fornece uma API que retorna informações sobre o clima pode ser chamada de "serviço de clima" e uma aplicação que calcula o imposto sobre uma venda pode ser chamada de "serviço de imposto". Para minimizar a confusão potencial, na seguinte seção usamos o termo "Serviço" com S maiúsculo para se referir à abstração do Kubernetes e "serviço" com s minúsculo para se referir ao sinônimo de aplicações.

Adicionando, Modificando e Removendo Serviços

Você pode adicionar, modificar e remover Serviços de um cluster usando tanto o Cloud Console quanto o Cloud SDK em seu ambiente local, em uma VM do Google Cloud, ou no Cloud Shell.

Um serviço é uma abstração que agrupa um conjunto de pods como um único recurso.

Adicionando, Modificando e Removendo Serviços com Cloud Console

Serviços do Kubernetes são adicionados através de deployments. No Cloud Console, selecione a opção Workloads no menu de navegação para exibir uma lista de deployments, conforme mostrado na Figura 8.31. Note a opção Deploy no menu horizontal no topo da página.

FIGURE 8.31 Deployment list along with a Deploy command to create new services

The screenshot shows the Google Cloud Console interface for managing workloads. At the top, there are navigation tabs for 'Workloads', 'REFRESH', 'DEPLOY' (highlighted in blue), and 'DELETE'. Below these are dropdown menus for 'Cluster' (set to 'Cluster') and 'Namespace' (set to 'Namespace'), and buttons for 'RESET' and 'SAVE'. A help section defines 'Workloads' as deployable units of computing. The main area has tabs for 'OVERVIEW' (selected) and 'COST OPTIMIZATION'. A filter bar shows 'Is system object: False' and a 'Filter workloads' input field. A table lists one deployment: 'nginx-1' (Status: OK, Type: Deployment, Pods: 3/3, Namespace: default, Cluster: standard-cluster-1). The table has columns for Name, Status, Type, Pods, Namespace, and Cluster.

Name	Status	Type	Pods	Namespace	Cluster
nginx-1	OK	Deployment	3/3	default	standard-cluster-1

Clique em Deploy para exibir o formulário de deployment, mostrado na Figura 8.32.

FIGURE 8.32 Form that lets you specify a new deployment for a service

Create a deployment

1 Container

Edit container

Existing container image
 New container image

Image path *
nginx:latest **SELECT**

Enter your image path, or choose from Google Container Registry. You can also try to deploy with official nginx image nginx:latest.

Environment variables

+ ADD ENVIRONMENT VARIABLE

Initial command

Overrides the default entrypoint of the container image.

CANCEL DONE

2 Configuration

ADD CONTAINER

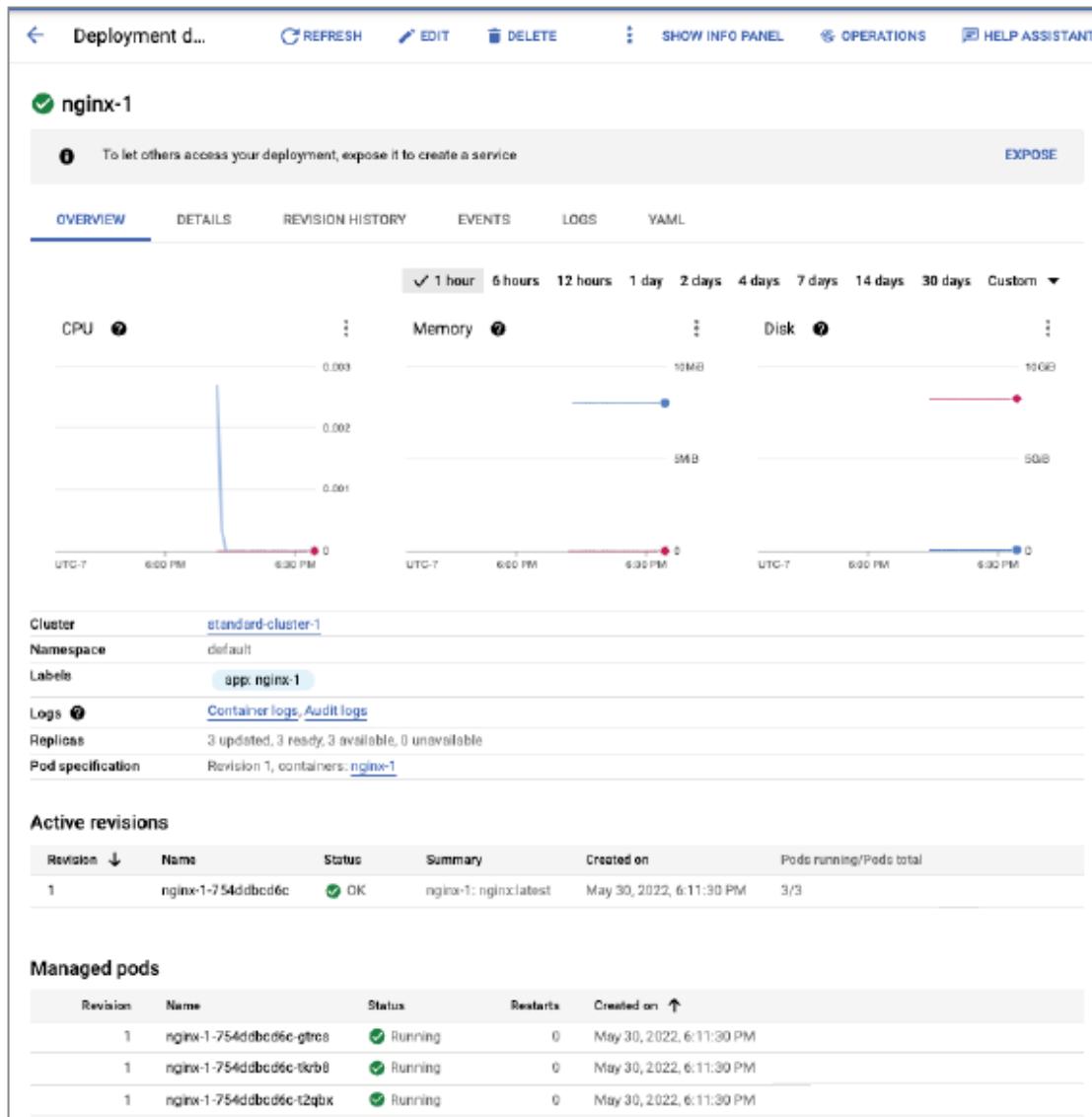
CONTINUE

No parâmetro Imagem do Contêiner, você pode especificar o nome de uma imagem ou selecionar uma do Google Container Repository. Para especificar um nome diretamente, especifique um caminho para a imagem usando uma URL como esta:

`gcr.io/google-samples/hello-app:2.0`

Você pode especificar rótulos, o comando inicial a ser executado e um nome para sua aplicação. Quando você clica no nome de um deployment, verá detalhes desse deployment, incluindo uma lista de Serviços, como o mostrado na Figura 8.33. Clicar no nome de um Serviço abre o formulário de Detalhe do Serviço, que inclui uma opção de Deletar no menu horizontal. A Figura 8.34 mostra a caixa de diálogo de exclusão.

FIGURE 8.33 Details of Services exposing a deployment

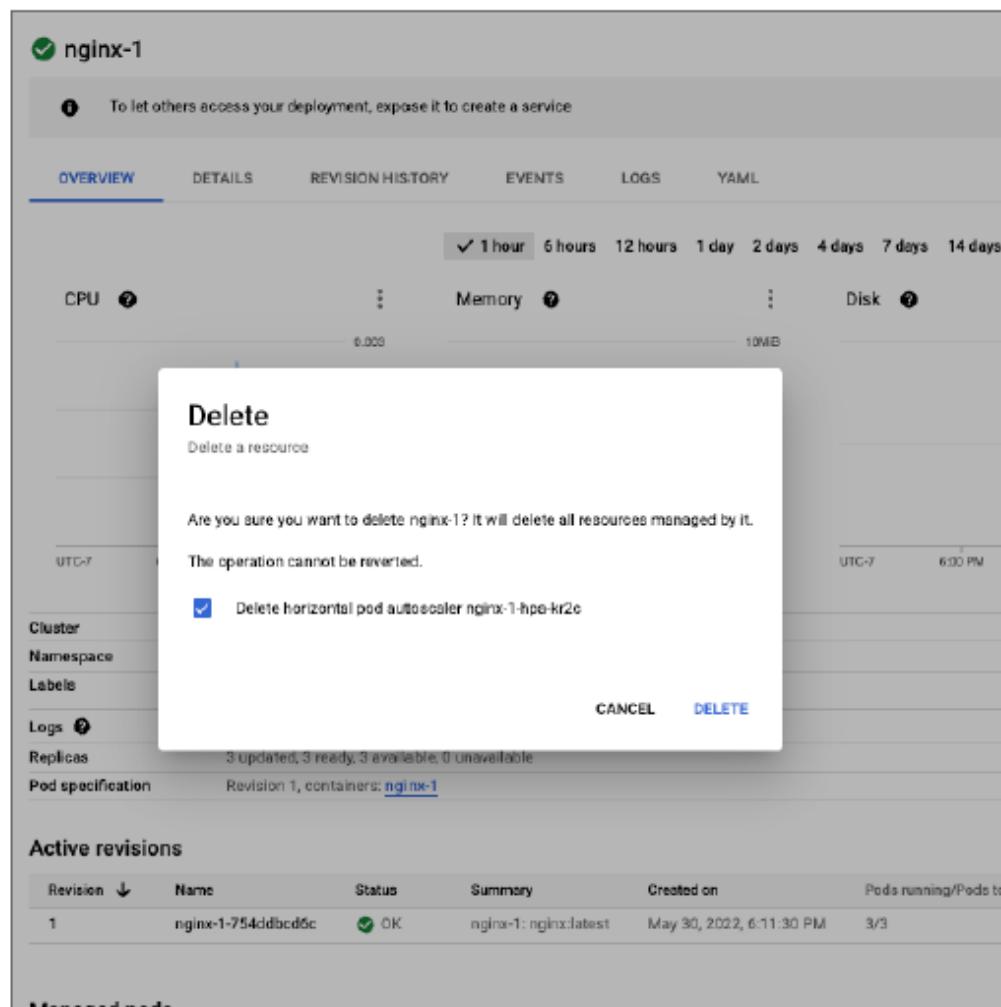


Adicionando, Modificando e Removendo Serviços com Cloud SDK e Cloud Shell

Use o comando `kubectl get services` para listar Serviços. Para adicionar um Serviço, use o comando `kubectl create deployment` para iniciar um Serviço. Por exemplo, para adicionar um Serviço chamado `hello-server` usando a aplicação de exemplo de mesmo nome fornecida pelo Google, use o seguinte comando:

```
kubectl create deployment hello-server --image=gcr.io/google/samples/hello-app:1.0 --port=8080
```

FIGURE 8.34 Navigate to the Service Details page to delete a service using the Delete option in the horizontal menu.



Este comando baixará e começará a executar a imagem encontrada no caminho gcr.io/google-samples/hello-app, versão 1. Ele estará acessível na porta 8080. Deployments precisam ser expostos para serem acessíveis a recursos fora do cluster. Isso pode ser configurado usando o comando expose, como mostrado aqui:

```
kubectl expose deployment hello-server --type="LoadBalancer"
```

Este comando expõe o Serviço fazendo com que um balanceador de carga atue como o ponto de contato para recursos externos contatarem o serviço. Para remover um Serviço, use o comando delete service, como mostrado aqui:

```
kubectl delete service hello-server
```

Criando Repositórios no Artifact Registry

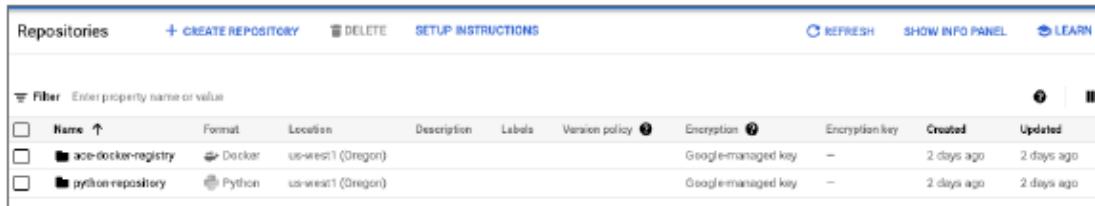
Artifact Registry é um serviço do Google Cloud para armazenar imagens de contêineres. Container Registry é um serviço usado no passado para gerenciar imagens, mas agora o Artifact Registry é o serviço recomendado para gerenciar imagens. Uma vez que você tenha criado um registro e enviado imagens para ele, você pode visualizar o

conteúdo do registro e detalhes da imagem usando o Cloud Console e o Cloud SDK e Cloud Shell.

Visualizando o Repositório de Imagens e Detalhes da Imagem com Cloud Console

No Cloud Console, selecione Artifact Registry no menu de navegação para exibir os registros de exemplo (veja Figura 8.35).

FIGURE 8.35 A listing of repositories in Artifact Registry



The screenshot shows the 'Repositories' page in the Google Cloud Artifact Registry. At the top, there are buttons for '+ CREATE REPOSITORY', 'DELETE', 'SETUP INSTRUCTIONS', 'REFRESH', 'SHOW INFO PANEL', and 'LEARN'. Below this is a search bar labeled 'Filter' with the placeholder 'Enter property name or value'. A table lists three repositories:

Name	Format	Location	Description	Labels	Version policy	Encryption	Encryption key	Created	Updated
ace-docker-registry	Docker	us-west1 (Oregon)				Google-managed key	—	2 days ago	2 days ago
python repository	Python	us-west1 (Oregon)				Google-managed key	—	2 days ago	2 days ago

Para criar um registro, clique no ícone + para exibir uma caixa de diálogo como a mostrada na Figura 8.36. Você pode ver que o Artifact Registry pode ter vários tipos de registros, incluindo um para Docker, Maven (um framework Java) e Python, entre outros. Dependendo do tipo de repositório que você criar, deve tomar etapas adicionais para configurar o repositório. O Artifact Registry fornece comandos detalhados para configurar o repositório. A Figura 8.37, por exemplo, mostra um comando para configurar um repositório Docker.

FIGURE 8.36 Creating a repository in Artifact Registry

Name *

Format

- Docker
- Maven
- npm
- Python
- Apt
- Yum
- Kubeflow Pipelines [PREVIEW](#)

Location type

- Region
- Multi-region

Region *

Description

Labels

[+ ADD LABEL](#)

Encryption

- Google-managed encryption key
No configuration required
- Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

CREATE **CANCEL**

FIGURE 8.37 Example instructions for configuring a Docker repository

My First Project ▾ Search artifact

Images for ace-docker-registry DELETE SETUP INSTRUCTIONS

us-west1-docker.pkg.dev > scenario-energy-335022 > ace-docker-registry

Filter Enter property name or value

Name	Created	Updated
No rows to display		

Setup instructions

Follow the steps below to configure your client to push and pull packages using this repository. You can also view more detailed instructions [here](#). For more information about working with artifacts in this repository, see the [documentation](#).

Initialize gcloud

The [Google Cloud SDK](#) is used to generate an access token when authenticating with Artifact Registry. Make sure that it is installed and initialized with [Application Default Credentials](#) before proceeding.

Configure Docker

Run the following command to configure `gcloud` as the credential helper for the Artifact Registry domain associated with this repository's location:

```
$ gcloud auth configure-docker us-west1-docker.pkg.dev
```

O Kubernetes Engine utiliza imagens de contêiner armazenadas em um Repositório Docker. O conteúdo do Repositório Docker pode ser visualizado em resumo e em detalhes usando tanto o Cloud Console quanto o SDK de linha de comando, incluindo no Cloud Shell.

Resumo

Neste capítulo, você aprendeu como realizar tarefas básicas de gerenciamento ao trabalhar com clusters Kubernetes, nós, pods e serviços. O capítulo também descreveu como listar o conteúdo dos repositórios de imagens de contêiner. Você aprendeu a fixar serviços no menu Cloud Console, visualizar o status dos clusters Kubernetes e ver detalhes do repositório de imagens e das imagens usando comandos gcloud. Este capítulo também descreveu como modificar e remover nós e pods. Você também viu os benefícios do autoscaling em um cenário do mundo real.

Tanto o Cloud Console quanto o Cloud SDK, incluindo o Cloud Shell, podem ser usados para adicionar, remover e modificar nós, pods e serviços. Ambos podem ser usados para revisar o conteúdo de um repositório de imagens. Alguns dos comandos mais úteis incluem gcloud container clusters create e gcloud container clusters resize. O comando kubectl é usado para modificar recursos do Kubernetes, como deployments e pods.

Essenciais para o Exame

Saiba como visualizar o status de um cluster Kubernetes. Use o Cloud Console para listar clusters e aprofundar nos clusters para ver detalhes do cluster, incluindo detalhes de nós, pods e contêineres. Conheça o comando gcloud container clusters e suas opções.

Entenda como adicionar, modificar e remover nós. Use o Cloud Console para modificar nós e saiba como adicionar e remover nós alterando deployments. Use o comando gcloud container clusters resize para adicionar e remover nós.

Entenda como adicionar, modificar e remover pods. Use o Cloud Console para modificar pods e para adicionar e remover pods alterando deployments. Use kubectl get deployments para listar deployments, kubectl scale deployment para modificar o número de deployments e kubectl autoscale deployment para habilitar o Autoscaling.

Entenda como adicionar, modificar e remover Serviços. Use o Cloud Console para modificar Serviços e adicionar e remover Serviços alterando deployments. Use kubectl create deployment para iniciar Serviços e kubectl expose deployment para tornar um Serviço acessível fora do cluster. Delete um serviço usando o comando kubectl delete service.

1. Você está executando vários microsserviços em um cluster Kubernetes. Você notou alguma degradação de desempenho. Após revisar alguns logs, você começa a pensar que o cluster pode estar configurado incorretamente e abre o Cloud Console para investigar. Como você vê os detalhes de um cluster específico?

- A. Digite o nome do cluster na barra de pesquisa.
 - B. Clique no nome do cluster.
 - C. Use o comando gcloud cluster details.
 - D. Nenhuma das opções acima.
2. Você está visualizando os detalhes de um cluster no Cloud Console e quer saber quantas vCPUs estão disponíveis no cluster. Onde você procuraria por essa informação?
- A. Seção de Node Pools da página de Detalhes dos Nós
 - B. Seção de Labels da página de Detalhes do Cluster
 - C. Linha de Resumo da página de Listagem dos Clusters
 - D. Nenhuma das opções acima.
3. Você foi designado para ajudar a diagnosticar problemas de desempenho com aplicações executando em vários clusters Kubernetes. A primeira coisa que você quer fazer é entender, em um nível alto, as características dos clusters. Qual comando você deveria usar?
- A. gcloud container list
 - B. gcloud container clusters list
 - C. gcloud clusters list
 - D. Nenhuma das opções acima.
4. Quando você tenta usar o comando kubectl pela primeira vez, você recebe uma mensagem de erro indicando que o recurso não pode ser encontrado ou você não pode se conectar ao cluster. Qual comando você usaria para tentar eliminar o erro?
- A. gcloud container clusters access
 - B. gcloud container clusters get-credentials
 - C. gcloud auth container
 - D. gcloud auth container clusters
5. Um engenheiro recentemente se juntou à sua equipe e não está ciente dos padrões da sua equipe para criar clusters e outros objetos Kubernetes. Em particular, o engenheiro não rotulou corretamente vários clusters. Como você modificaria os rótulos do cluster pelo Cloud Console?

- A. Clique no botão Conectar.
 - B. Clique na opção de menu Deploy.
 - C. Clique na opção de menu Editar.
 - D. Digite os novos rótulos na seção de Labels.
6. Você recebe uma página no meio da noite informando que vários serviços rodando em um cluster Kubernetes têm alta latência ao responder a solicitações de API. Você revisa os dados de monitoramento e determina que não há recursos suficientes no cluster para acompanhar a carga. Você decide adicionar mais seis VMs ao cluster. Quais parâmetros você precisará especificar quando emitir o comando de redimensionamento do cluster?
- A. Tamanho do cluster
 - B. Nome do cluster
 - C. Nome do pool de nós
 - D. Todas as opções acima.
7. Você quer modificar o número de pods em um cluster. Qual é a melhor maneira de fazer isso?
- A. Modificar os pods diretamente
 - B. Modificar os deployments
 - C. Modificar os pools de nós diretamente
 - D. Modificar os nós
8. Você quer ver uma lista de deployments. Qual opção do menu de navegação do Kubernetes Engine você selecionaria?
- A. Clusters
 - B. Armazenamento
 - C. Workloads
 - D. Deployments
9. Quais ações estão disponíveis no menu Ações ao visualizar os detalhes do deployment?
- A. Escalar e Autoscalar apenas
 - B. Autoscalar, Expor e Atualização Gradual
 - C. Adicionar, Modificar e Deletar
 - D. Nenhuma das opções acima

10. Qual é o comando para listar deployments a partir da linha de comando?

- A. gcloud container clusters list-deployments
- B. gcloud container clusters list
- C. kubectl get deployments
- D. kubectl deployments list

11. Onde você pode visualizar uma lista de aplicações ao usar o Cloud Console?

- A. Na página de Detalhes do Deployment
- B. Na página de Detalhes do Contêiner
- C. Na página de Detalhes do Cluster
- D. Nenhuma das opções acima

12. Qual comando kubectl é usado para criar um deployment?

- A. run
- B. start
- C. initiate
- D. deploy

13. Você está dando suporte a engenheiros de aprendizado de máquina que estão testando uma série de classificadores. Eles têm cinco classificadores, chamados ml-classifier-1, ml-classifier-2, etc. Eles descobriram que o ml-classifier-3 não está funcionando conforme o esperado e gostariam que ele fosse removido do cluster. O que você faria para deletar um serviço chamado ml-classifier-3?

- A. Execute o comando kubectl delete service ml-classifier-3.
- B. Execute o comando kubectl delete ml-classifier-3.
- C. Execute o comando gcloud service delete ml-classifier-3.
- D. Execute o comando gcloud container service delete ml-classifier-3.

14. Qual serviço é responsável por gerenciar imagens de contêineres?

- A. Kubernetes Engine
- B. Compute Engine
- C. Artifact Registry
- D. Container Engine

15. Qual comando é usado para listar imagens de contêineres na linha de comando?

- A. gcloud container images list
 - B. gcloud container list images
 - C. kubectl list container images
 - D. kubectl container list images
16. Um projetista de data warehouse quer implantar um processo de extração, transformação e carga para o Kubernetes. O designer forneceu uma lista de bibliotecas que devem ser instaladas, incluindo drivers para GPUs. Você tem várias imagens de contêineres que acha que podem atender aos requisitos. Como você poderia obter uma descrição detalhada de cada um desses contêineres?
- A. Execute o comando gcloud container images list details.
 - B. Execute o comando gcloud container images describe.
 - C. Execute o comando gcloud image describe.
 - D. Execute o comando gcloud container describe.
17. Você acabou de criar um deployment e quer que aplicações fora do cluster tenham acesso aos pods fornecidos pelo deployment. O que você precisa fazer com o deployment?
- A. Dar a ele um endereço IP público.
 - B. Emitir um comando kubectl expose deployment.
 - C. Emitir um comando gcloud expose deployment.
 - D. Nada; torná-lo acessível deve ser feito no nível do cluster.
18. Você implantou uma aplicação em um cluster Kubernetes que processa dados de sensores de uma frota de veículos de entrega. O volume de dados de entrada depende do número de veículos fazendo entregas. O número de veículos fazendo entregas depende do número de pedidos dos clientes. Os pedidos dos clientes são altos durante as horas do dia, temporadas de férias e quando campanhas publicitárias importantes são realizadas. Você quer garantir que tenha nodes suficientes rodando para lidar com a carga, mas quer manter seus custos baixos. Como você configuraria seu cluster Kubernetes?
- A. Implante tantos nodes quanto seu orçamento permitir.
 - B. Habilite o Autoscaling.
 - C. Monitore a utilização da CPU, disco e rede e adicione nodes conforme necessário.
 - D. Escreva um script para executar comandos gcloud para adicionar e remover nodes quando os picos geralmente começam e terminam, respectivamente.

19. Ao usar o Kubernetes Engine, o que um engenheiro de nuvem pode precisar configurar?

- A. Nodes, pods, serviços e clusters apenas
- B. Nodes, pods, serviços, clusters e imagens de contêineres
- C. Nodes, pods, clusters e imagens de contêineres apenas
- D. Pods, serviços, clusters e imagens de contêineres apenas

20. Quais parâmetros de um deployment podem ser configurados na página Criar Deployment no Cloud Console?

- A. Imagem do contêiner
- B. Nome do cluster
- C. Nome da aplicação
- D. Todas as opções acima

Capítulo 9

Computação com Cloud Run e App Engine

ESTE CAPÍTULO COBRE O SEGUINTE OBJETIVO DO EXAME DE CERTIFICAÇÃO DO GOOGLE ASSOCIATE CLOUD ENGINEER:

3.3 Implantando e implementando recursos do Cloud Run e Cloud Functions

Este capítulo descreve como implantar serviços conteinerizados usando o Cloud Run e o App Engine. O Cloud Run, uma alternativa ao App Engine para executar contêineres, é um serviço gerenciado e sem servidor. O App Engine não está mais incluído no Guia do Exame de Certificação do Associate Cloud Engineer; no entanto, ainda está incluído neste capítulo porque ainda é uma opção no Google Cloud e os engenheiros de nuvem devem ser capazes de suportar os serviços do App Engine, mesmo que não sejam feitas perguntas sobre o App Engine em um exame.

O Cloud Run é projetado para suportar aplicações conteinerizadas altamente escaláveis, escritas em qualquer linguagem. O Cloud Run integra-se com ferramentas de desenvolvimento, como o Cloud Build, Artifact Registry e Docker.

O App Engine é uma oferta de plataforma como serviço (PaaS) do Google Cloud que permite aos desenvolvedores trabalharem dentro de um conjunto de frameworks específicos de linguagem e implantarem aplicações escaláveis com apenas uma atenção mínima às preocupações de escalabilidade.

Visão Geral do Cloud Run

O Cloud Run é um serviço gerenciado e sem servidor para executar aplicações conteinerizadas. Ao contrário do App Engine Standard e do Cloud Function, você não está restrito a usar um conjunto limitado de linguagens de programação. O Cloud Run suporta qualquer aplicação que possa ser executada em um contêiner. Uma vantagem de usar o Cloud Run é que você não precisa gerenciar a infraestrutura, como máquinas virtuais. O Cloud Run suporta duas maneiras de executar código: como um serviço e como um trabalho.

Os serviços do Cloud Run são usados quando o seu código é utilizado para responder a solicitações da web ou eventos. Por exemplo, uma API que retorna dados de um banco de dados Cloud SQL poderia ser implementada usando contêineres e executada como um serviço no Cloud Run.

Os trabalhos do Cloud Run são usados quando o código executa até que uma carga de trabalho esteja completa. Por exemplo, se você precisasse transformar um conjunto de arquivos armazenados no Cloud Storage e depois carregá-lo no Cloud SQL, você poderia executar a aplicação em um contêiner usando os trabalhos do Cloud Run.

Serviços do Cloud Run

Os serviços do Cloud Run são bem adequados para aplicações web, microsserviços, APIs e processamento de dados em stream.

Os serviços do Cloud Run são projetados para escutar um endpoint HTTPS e responder a solicitações feitas a esse endpoint. Cada serviço do Cloud Run possui um endpoint em um subdomínio único do domínio run.app, e domínios personalizados também podem ser usados. Os endpoints podem escalar até 1.000 instâncias de contêiner com quotas padrão; você pode solicitar uma quota maior, se necessário.

Você também pode especificar um número máximo de instâncias de contêineres para executar, caso queira limitar o número de contêineres e, portanto, o custo de executá-los. Além de fornecer recursos escaláveis para suportar o tráfego para o endpoint, o Cloud Run gerencia TLS. Você pode usar WebSockets, HTTP/2 (end-to-end) e gRPC com esses endpoints. Com o Cloud Run, você implanta versões imutáveis de um serviço. Para fazer uma atualização em um serviço, você criaria uma nova imagem de contêiner e a implantaria como uma nova versão. Você pode executar várias revisões do mesmo serviço no Cloud Run. Além disso, você pode rotear o tráfego entre diferentes revisões. Isso é útil quando você lança uma nova versão e deseja enviar uma pequena quantidade de tráfego para a versão mais recente, para que possa monitorar qualquer problema antes de disponibilizar a última versão para todos os usuários. Se você descobrir problemas com uma revisão, pode reverter e direcionar o tráfego para uma revisão anterior, mais estável.

Os serviços do Cloud Run são implantados de forma privada por padrão e requerem autenticação para acesso. Você pode controlar o acesso aos serviços das seguintes maneiras:

- Com uma política de Cloud IAM
- Usando configurações de ingresso

- Permitindo apenas usuários autenticados com o Cloud Identity Aware Proxy (IAP)

Com políticas de Cloud IAM, você pode atribuir um papel a um grupo de usuários para que o grupo tenha as permissões especificadas no papel. Por exemplo, para tornar um serviço publicamente acessível, você pode permitir o acesso a usuários não autenticados. Você pode querer conceder a um grupo de desenvolvedores permissões para criar novas versões de um serviço, e você pode fazer isso atribuindo o papel de desenvolvedor do run.

Você também pode controlar o acesso no nível da rede. Por padrão, um endpoint do Cloud Run é acessível de qualquer lugar na Internet usando o subdomínio run.app ou um domínio personalizado que você definir. Além de usar papéis do IAM para controlar o acesso a um serviço, você pode controlar o tráfego de rede para o endpoint especificando uma configuração de ingresso. As opções de configuração de ingresso incluem:

- Interno, que é o mais restritivo e permite apenas tráfego de平衡adores de carga HTTP(S) internos, recursos dentro do perímetro de Controles de Serviço da VPC, redes VPC no mesmo projeto ou perímetro de Controles de Serviço da VPC, bem como serviços Eventarc, Cloud Pub/Sub e Cloud Workflow no mesmo projeto ou perímetro de controle de serviço
- Interno e Cloud Load Balancing, que inclui tráfego permitido pela configuração Interna, junto com平衡adores de carga HTTP(S) externos
- Todos, que é o menos restritivo e permite todas as solicitações enviadas ao endpoint do serviço

O Cloud IAP é um serviço de segurança que protege os serviços permitindo apenas tráfego para os serviços que vêm de proxies. Quando um usuário tenta acessar um serviço do Cloud Run protegido pelo Cloud IAM, ele é primeiro sujeito à autenticação e autorização pelo IAP.

Trabalhos do Cloud Run

Trabalhos do Cloud Run são programas ou scripts que executam por um período de tempo enquanto completam uma tarefa e então param. Por exemplo, você poderia usar trabalhos do Cloud Run para executar um script para validar arquivos carregados em um bucket do Cloud Storage e então importar dados desses arquivos para um banco de dados Cloud SQL. Ao contrário de um serviço, que continua aceitando solicitações para realizar tarefas, trabalhos realizam uma tarefa e terminam. Trabalhos do Cloud Run podem ser agendados para serem executados em horários regulares. Eles também podem ser trabalhos em array, que podem ser paralelizados. O exemplo de processamento de arquivo mencionado é um bom exemplo de um trabalho que pode ser paralelizado. Em vez de processar cada arquivo um de cada vez, vários contêineres podem ser iniciados para que vários arquivos possam ser processados simultaneamente.

Criando um Serviço Cloud Run

Você pode criar um serviço Cloud Run usando o console, o Cloud SDK ou programaticamente usando a API. Nesta seção, vamos revisar como criar um serviço Cloud Run usando o console.

No console da nuvem, navegue até a página do Cloud Run e selecione a opção para criar um serviço. A página de Criação de Serviço, mostrada na Figura 9.1, será aberta.

FIGURE 9.1 The form for creating a Cloud Run service

The screenshot shows the 'Create service' page for Cloud Run. At the top, there's a header with 'Cloud Run' and a back arrow. Below it, a note says: 'A service exposes a unique endpoint and automatically scales the underlying infrastructure to handle incoming requests. Service name and region cannot be changed later.' There are two main deployment options:

- Deploy one revision from an existing container image:
Container image URL: `us-docker.pkg.dev/cloudbuild/container/hello` (with a 'SELECT' button) and a 'TEST WITH A SAMPLE CONTAINER' link.
- Continuously deploy new revisions from a source repository.

Below these, configuration fields include:

- Service name: `hello`
- Region: `us-west1 (Oregon)` (with a 'How to pick a region?' link)
- CPU allocation and pricing:
 - CPU is only allocated during request processing: 'You are charged per request and only when the container instance processes a request.'
 - CPU is always allocated: 'You are charged for the entire lifecycle of the container instance.'
- Autoscaling:
 - Minimum number of instances: `0`
 - Maximum number of instances: `100`

Set to 1 to reduce cold starts. [Learn more](#)

To the right, a sidebar titled 'Cloud Run pricing' lists the 'Free tier' benefits:

- First 180,000 vCPU-seconds/month
- First 360,000 GiB-seconds/month
- 2 million requests/month

At the bottom of the sidebar is a link: '→ Check paid tiers details'.

Nesta página, você especificará uma URL de imagem de contêiner. Você digita uma URL ou seleciona uma imagem do Container Registry ou do Artifact Registry. Por padrão, você implantará uma revisão, mas pode selecionar a opção para implantar continuamente novas versões conforme o repositório fonte é atualizado.

Você também especificará um nome de serviço e escolherá uma região para executar seu serviço. Você tem a opção de pagar apenas pelo tempo em que os recursos de CPU são alocados para processar uma solicitação ou pagar por recursos de CPU que estão sempre alocados. Você também pode especificar um número mínimo e máximo de instâncias.

A Figura 9.2 mostra como podemos especificar uma regra de ingresso. As opções de configuração de ingresso foram descritas acima. Você também pode alterar o requisito

padrão para autenticação para permitir acesso não autenticado. O acesso não autenticado é tipicamente usado para sites ou APIs públicas.

FIGURE 9.2 When creating a Cloud Run service, we can choose one of three ingress options.



Em seguida, você pode especificar opções de configuração adicionais para contêineres, conexões e segurança.

Para contêineres (Figura 9.3), você pode especificar uma porta, um comando de contêiner e argumentos de contêiner. Você também pode configurar a quantidade de memória e número de CPUs. Atualmente, a memória máxima é de 32 GB em prévia e 16 GB no lançamento geral. Até 8 CPUs são atualmente suportadas em prévia e até 4 no lançamento geral. (Serviços em prévia não são cobertos pelo SLA do Google Cloud, mas serviços no lançamento geral são cobertos por SLAs.)

Por padrão, as solicitações irão expirar após 5 minutos, mas você pode especificar um período mais curto ou mais longo, variando de 1 a 60 minutos.

O Cloud Run tem dois ambientes de execução: primeira geração e segunda geração. A segunda geração suporta recursos como acesso ao sistema de arquivos e desempenho mais rápido. Por padrão, o Cloud Run escolherá um ambiente para você. Você também pode especificar variáveis de ambiente para o contêiner e referenciar segredos no contêiner.

Na aba de Conexões (Figura 9.4), você pode indicar se deseja usar HTTP/2 end-to-end, que suporta serviços de streaming gRPC, e se deseja suportar afinidade de sessão. Afinidade de sessão irá rotear solicitações de um cliente para o mesmo contêiner, se possível. Você também pode especificar uma conexão Cloud SQL para serviços que usam um banco de dados Cloud SQL. Você também pode criar um Conector VPC para usar o Acesso VPC Serverless para conectar seu serviço Cloud Run a outros recursos em seu VPC, como instâncias do Compute Engine ou um cache Cloud Memorystore.

FIGURE 9.3 Configuring container parameters in a Cloud Run service

Container, Connections, Security

CONTAINER CONNECTIONS SECURITY

General

Container port 8080

Requests will be sent to the container on this port. We recommend listening on \$PORT instead of this specific number.

Container command

Leave blank to use the entry point command defined in the container image.

Container arguments

Arguments passed to the entry point command.

Capacity

Memory 512 MiB **CPU** 1

Memory to allocate to each container instance. Number of vCPUs allocated to each container instance.

Request timeout 300 seconds

Time within which a response must be returned (maximum 3600 seconds).

Maximum requests per container 80

The maximum number of concurrent requests that can reach each container instance. [What is concurrency?](#)

Execution environment

The execution environment your container runs in. [Learn More](#)

Default
Cloud Run will select a suitable execution environment for you.

First generation

Second generation PREVIEW
File system access, full Linux compatibility, faster performance.

Environment variables

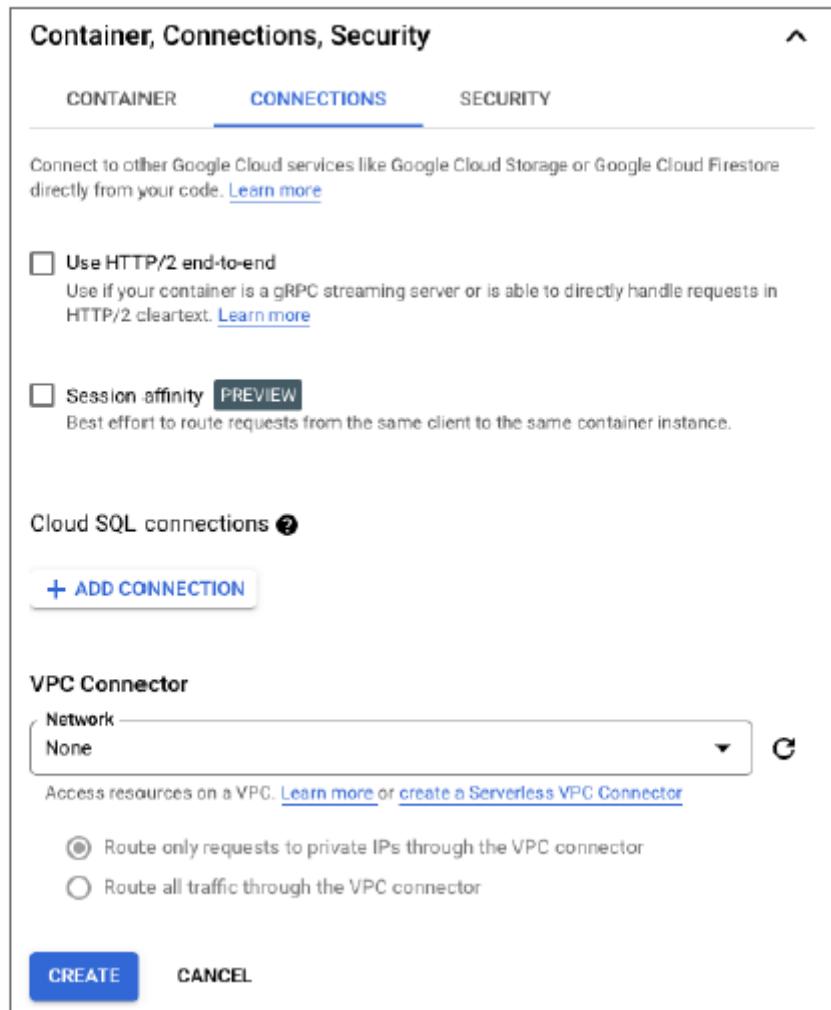
[+ ADD VARIABLE](#)

Secrets ?

[REFERENCE A SECRET](#)

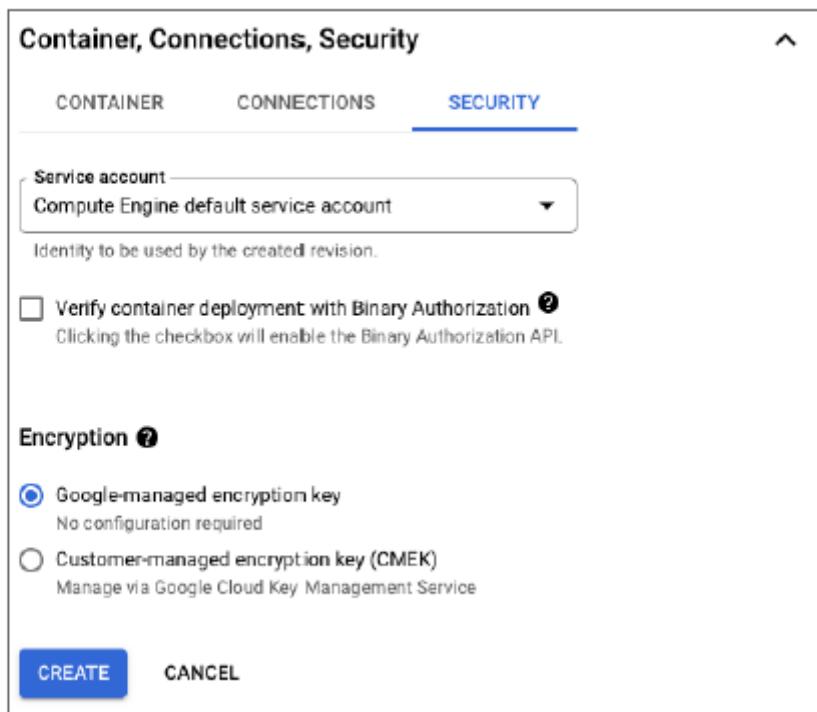
CREATE **CANCEL**

FIGURE 9.4 Configuring connection parameters in a Cloud Run service



Na aba de Segurança (Figura 9.5), você pode especificar uma conta de serviço para usar com este serviço. Você também pode exigir Autorização Binária antes de implantar um contêiner. Autorização Binária é um serviço que verifica se os contêineres atendem aos requisitos especificados em uma política que governa a implantação de contêineres no Cloud Run e no Kubernetes Engine, entre outros serviços. Você também pode especificar se deseja usar chaves de criptografia gerenciadas pelo Google ou pelo cliente.

FIGURE 9.5 Configuring security parameters in a Cloud Run service



Criando um Trabalho Cloud Run

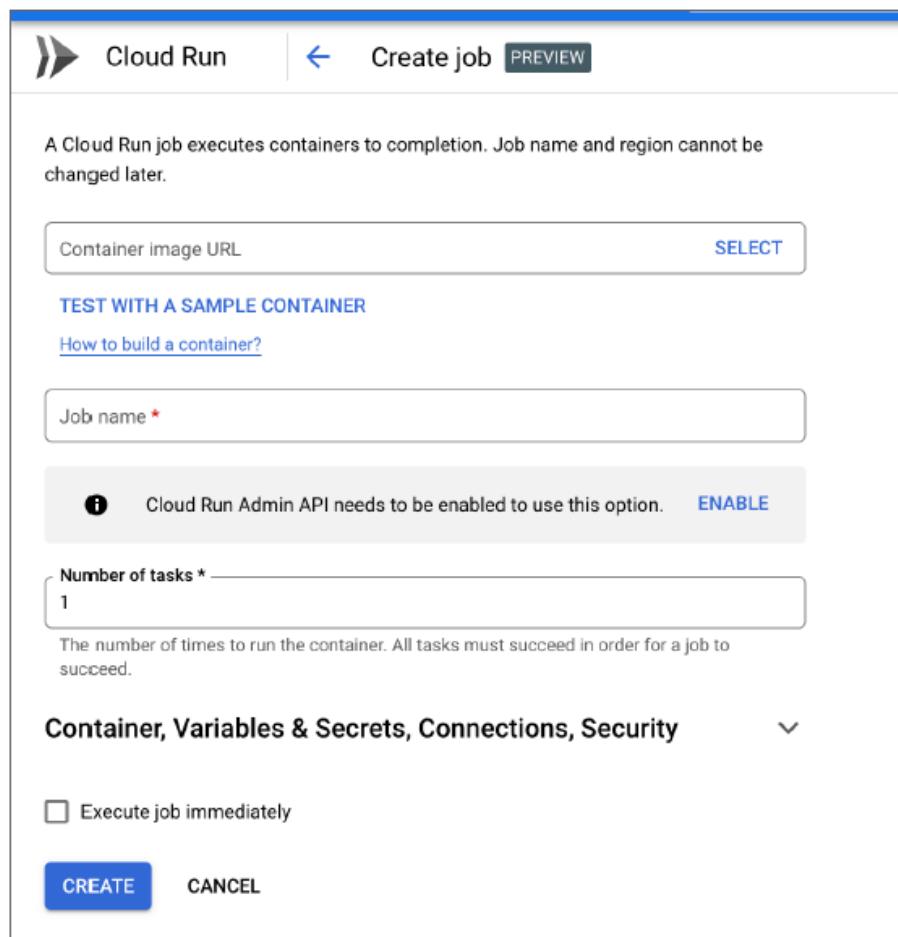
Criar um trabalho no Cloud Run é semelhante a criar um serviço. Na página Cloud Run no console da nuvem, selecione a aba Trabalhos e Criar um Trabalho para abrir um formulário semelhante à Figura 9.6. Assim como ao criar um serviço, você especificará uma URL de imagem de contêiner e região. Você também fornecerá um nome para o trabalho e o número de vezes que deseja executar o contêiner; o padrão é uma vez.

Na aba Geral, você pode especificar parâmetros de configuração de contêiner (Figura 9.7). Alguns parâmetros, como Comando do Contêiner, Argumentos do Contêiner, Memória e CPU são semelhantes ao que você viu ao configurar um serviço Cloud Run. Além disso, você pode especificar o número de tentativas de tarefas falhadas e um parâmetro de paralelismo para controlar o número de tarefas concorrentes. Há também uma opção para executar um trabalho imediatamente.

Na aba Variáveis & Segredos (Figura 9.8), você pode especificar variáveis de ambiente e referências a segredos armazenados. Assim como nos serviços Cloud Run, você pode especificar conexões Cloud SQL e um conector VPC na aba Conexões (Figura 9.9). Na aba Segurança (Figura 9.10), você pode especificar uma conta de serviço para o serviço e gerenciamento de chave de criptografia.

Antes do lançamento do Cloud Run, os desenvolvedores muitas vezes optavam por executar seus serviços no App Engine.

FIGURE 9.6 Creating a Cloud Run job



Componentes do App Engine

O App Engine está disponível em uma versão Padrão e uma versão Flexível. Aplicações App Engine Padrão consistem em quatro componentes:

- Aplicação
- Serviço
- Versão
- Instância

Uma aplicação App Engine é um recurso de alto nível criado em um projeto; ou seja, cada projeto pode ter uma aplicação App Engine. Todos os recursos associados a um app App Engine são criados na região especificada quando o app é criado.

FIGURE 9.7 Configuring container parameters for a Cloud Run job

Container, Variables & Secrets, Connections, Security ^

GENERAL VARIABLES & SECRETS CONNECTIONS

Container command
Leave blank to use the entry point command defined in the container image.

Container arguments
Arguments passed to the entry point command.

Task capacity

Memory — 512 MiB CPU — 1
Memory to allocate to each container instance. Number of vCPUs allocated to each container instance.

Task timeout * 600 seconds
The maximum amount of time an instance can run for (maximum 3600 seconds).

Number of retries per failed task * 0

Parallelism
The maximum number of tasks running at the same time.

Run as many tasks concurrently as possible
For 1 CPU Cloud Run can run up to 100 tasks at a time.

Limit the number of concurrent tasks
Use this option to limit the number of concurrent requests to backing resources such as databases or file systems.

Execute job immediately

FIGURE 9.8 Configuring variables and secrets for a Cloud Run job

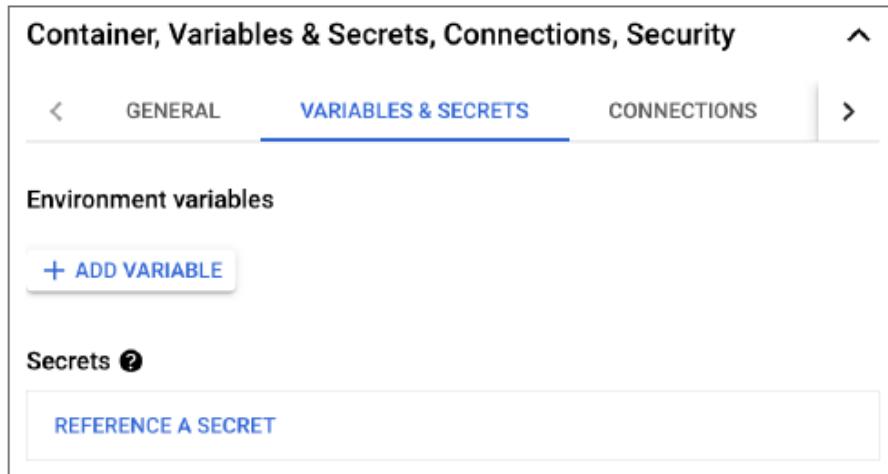


FIGURE 9.9 Configuring connection parameters for a Cloud Run job

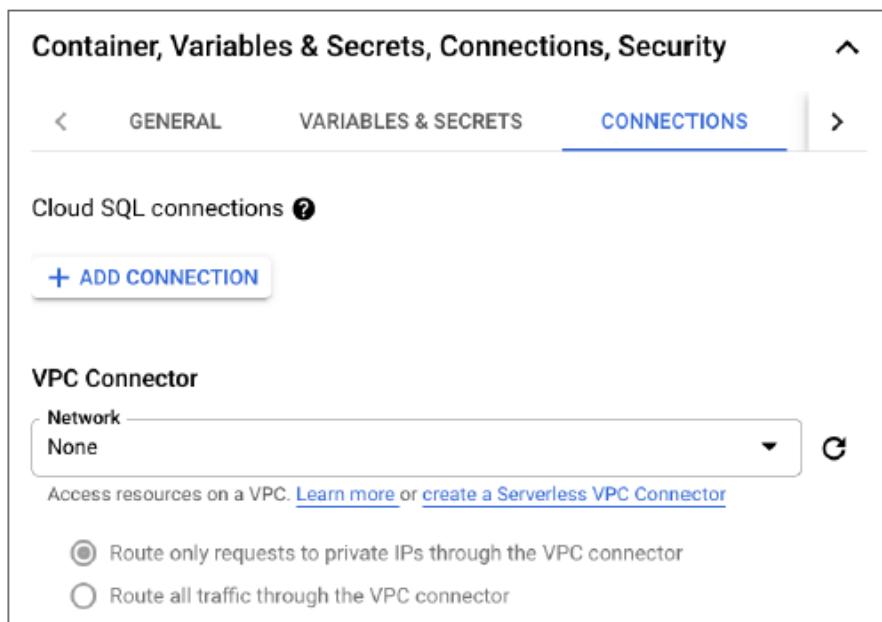
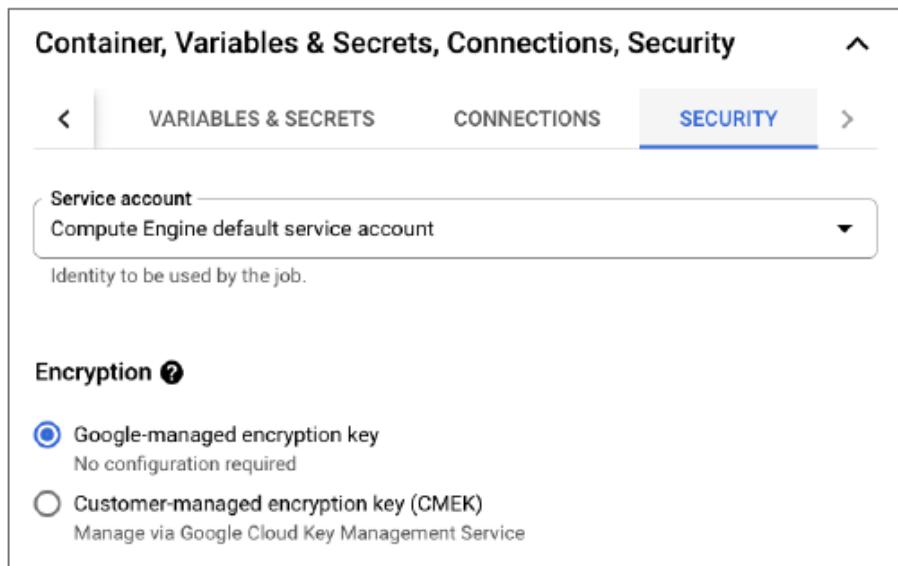


FIGURE 9.10 Configuring security parameters for a Cloud Run job



Apps têm pelo menos um serviço, que é o código executado no ambiente App Engine. Como várias versões da base de código de uma aplicação podem existir, o App Engine suporta versionamento de apps. Um serviço pode ter várias versões, e estas geralmente são ligeiramente diferentes, com versões mais novas incorporando novos recursos, correções de bugs e outras mudanças em relação às versões anteriores. Quando uma versão é executada, ela cria uma instância do app.

Serviços são tipicamente estruturados para realizar uma única função, com aplicações complexas compostas por vários serviços, conhecidos como microsserviços. Um microsserviço pode lidar com solicitações de API para acesso a dados, enquanto outro microsserviço realiza autenticação e um terceiro registra dados para fins de faturamento.

Os serviços são definidos pelo seu código-fonte e pelo seu arquivo de configuração. A combinação desses arquivos constitui uma versão do aplicativo. Se você alterar ligeiramente o código-fonte ou o arquivo de configuração, outra versão será criada. Dessa forma, você pode manter várias versões do seu aplicativo ao mesmo tempo, o que é especialmente útil para testar novos recursos em um pequeno número de usuários antes de disponibilizar a mudança para todos os usuários. Se bugs ou outros problemas ocorrerem com uma versão, você pode facilmente voltar para uma versão anterior. Outra vantagem de manter várias versões é que elas permitem migrar e dividir o tráfego, o que descreveremos com mais detalhes mais adiante no capítulo.

Implantando um Aplicativo App Engine

O exame de certificação Google Associate Cloud Engineer não exige que os engenheiros escrevam um aplicativo, mas espera-se que saibam como implantar um. Nesta seção, você baixará um exemplo de Hello World do Google e usará como um aplicativo de amostra que você irá implantar. O app é escrito em Python, então você usará o runtime de Python no App Engine.

Implantando um App Usando Cloud Shell e SDK

Primeiro, você trabalhará em uma janela de terminal usando o Cloud Shell, que você pode iniciar a partir do console clicando no ícone do Cloud Shell. Certifique-se de que o gcloud esteja configurado para trabalhar com o App Engine usando o seguinte comando:

```
gcloud components install app-engine-python
```

Este comando instalará ou atualizará a biblioteca App Engine Python conforme necessário. Se a biblioteca estiver atualizada, você receberá uma mensagem dizendo isso.

Quando você abrir o Cloud Shell, você pode ter um diretório chamado python-docs-samples. Este contém uma série de aplicações de exemplo, incluindo o app Hello World que usaremos. Se você não ver este diretório, você pode baixar o app Hello World do Google usando isso:

```
git clone https://github.com/GoogleCloudPlatform/python-docs-samples
```

Em seguida, mude seu diretório de trabalho para o diretório com o app Hello World, usando o seguinte:

```
cd python-docs-samples/appengine/standard_python3/hello_world
```

Se você listar os arquivos no diretório, verá cinco arquivos:

- app.yaml
- main.py
- main_test.py
- requirements.txt
- requirements-test.txt

Aqui você está principalmente preocupado com o arquivo app.yaml (Figura 9.11).

FIGURE 9.11 The contents of an `app.yaml` file for a Python 3 application

```
1 # Copyright 2021 Google LLC
2 #
3 # Licensed under the Apache License, Version 2.0 (the "License");
4 # you may not use this file except in compliance with the License.
5 # You may obtain a copy of the License at
6 #
7 #     http://www.apache.org/licenses/LICENSE-2.0
8 #
9 # Unless required by applicable law or agreed to in writing, software
10 # distributed under the License is distributed on an "AS IS" BASIS,
11 # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12 # See the License for the specific language governing permissions and
13 # limitations under the License.
14
15 runtime: python27
16 api_version: 1
17 threadsafe: true
18
19 handlers:
20 - url: /*
21   script: main.app
```

Neste exemplo, o arquivo de configuração do aplicativo especifica a versão do Python a ser usada. Dependendo da versão do Python que você está usando, o arquivo `app.yaml` também pode conter a versão da API que você está implantando, um parâmetro Python chamado `threadsafe` e variáveis de ambiente.

Para implantar seu aplicativo, você pode usar o seguinte comando:

```
gcloud app deploy app.yaml
```

No entanto, `app.yaml` é o padrão, então se você estiver usando esse nome para o arquivo, não precisa especificar `app.yaml` no comando de implantação.

Este comando deve ser executado a partir do diretório com o arquivo `app.yaml`. O comando `gcloud app deploy` possui alguns parâmetros opcionais:

`--version`, para especificar um ID de versão personalizado

`--project`, para especificar o ID do projeto a ser usado para este aplicativo

`--no-promote`, para implantar o aplicativo sem direcionar o tráfego para ele

Você pode ver a saída do programa Hello World navegando em um navegador para a URL do seu projeto, como <https://gcpace-project.appspot.com>. A URL do projeto é o nome do projeto seguido por `.appspot.com`.

Você também pode atribuir um domínio personalizado se preferir não usar um URL `appspot.com`. Você pode fazer isso a partir da função Adicionar Novo Domínio Personalizado na página de Configurações do App Engine.

Você pode parar de servir versões usando o comando `gcloud app versions stop` e passando uma lista de versões para parar. Por exemplo, para parar de servir versões nomeadas `v1` e `v2`, use o seguinte:

```
gcloud app versions stop v1 v2
```

Escalando Aplicações App Engine

Instâncias são criadas para executar um aplicativo em um servidor gerenciado pelo App Engine. O App Engine pode adicionar ou remover automaticamente instâncias conforme necessário, com base na carga. Quando as instâncias são escaladas com base na carga, elas são chamadas de instâncias dinâmicas. Essas instâncias dinâmicas ajudam a otimizar seus custos, desligando quando a demanda é baixa.

Alternativamente, você pode configurar suas instâncias para serem residentes ou em execução o tempo todo. Essas são otimizadas para desempenho, para que os usuários esperem menos tempo enquanto uma instância é iniciada.

Sua configuração determina se uma instância é residente ou dinâmica. Se você configurar o autoescalamento ou escalamento básico, então as instâncias serão dinâmicas. Se você configurar o escalamento manual, então suas instâncias serão residentes.

Para especificar o escalonamento automático, adicione uma seção ao app.yaml que inclui o termo `automatic_scaling` seguido por pares de chave-valor de opções de configuração. Estes incluem:

- `target_cpu_utilization`
- `target_throughput_utilization`
- `max_concurrent_requests`
- `max_instances`
- `min_instances`
- `max_pending_latency`
- `min_pending_latency`

`target_cpu_utilization` Especifica a utilização máxima de CPU que ocorre antes que instâncias adicionais sejam iniciadas.

`target_throughput_utilization` Especifica o número máximo de solicitações simultâneas antes que instâncias adicionais sejam iniciadas. Isso é especificado como um número entre 0.5 e 0.95.

`max_concurrent_requests` Especifica as solicitações simultâneas máximas que uma instância pode aceitar antes de iniciar uma nova instância. O padrão é 10; o máximo é 80.

`max_instances` e `min_instances` Indica o intervalo do número de instâncias que podem ser executadas para este aplicativo.

`max_pending_latency` e `min_pending_latency` Indica o tempo máximo e mínimo que uma solicitação esperará na fila para ser processada.

Você também pode usar o escalonamento básico para habilitar o escalonamento automático. Os únicos parâmetros para o escalonamento básico são `idle_timeout` e `max_instances`.

Mundo Real

Microserviços vs. Aplicações Monolíticas

Aplicações escaláveis são frequentemente escritas como coleções de microserviços. Isso nem sempre foi o caso. No passado, muitas aplicações eram monolíticas, ou projetadas para incluir toda a funcionalidade em um único programa compilado ou script. Isso pode parecer uma maneira mais simples e fácil de gerenciar aplicações, mas na prática cria mais problemas do que resolve:

- Qualquer alteração na aplicação requer a redistribuição de toda a aplicação, o que pode levar mais tempo do que a distribuição de microserviços. Desenvolvedores tendiam a agrupar mudanças antes de lançá-las.
- Se um lançamento agrupado tivesse um bug em uma mudança de funcionalidade, então todas as mudanças de funcionalidade seriam revertidas quando a aplicação monolítica fosse revertida.
- Era difícil coordenar mudanças quando equipes de desenvolvedores tinham que trabalhar com um único arquivo ou um pequeno número de arquivos de código-fonte.

Microserviços dividem o código da aplicação em aplicações de função única, permitindo que os desenvolvedores alterem um serviço e o distribuam sem impactar outros serviços. Ferramentas de gerenciamento de código-fonte, como Git, facilitam para vários desenvolvedores contribuir com componentes de um sistema maior, coordenando mudanças em arquivos de código-fonte. Este código de função única e a fácil integração com outro código promovem atualizações mais frequentes e a capacidade de testar novas versões antes de distribuí-las para todos os usuários de uma vez.

Dividindo o Tráfego Entre Versões do App Engine

Se você tiver mais de uma versão de uma aplicação em execução, pode dividir o tráfego entre as versões. O App Engine oferece três maneiras de dividir o tráfego: por endereço IP, por cookie HTTP e por seleção aleatória. A divisão por endereço IP proporciona alguma aderência, de modo que um cliente é sempre roteado para o mesmo segmento, pelo menos enquanto o endereço IP não mudar. Cookies HTTP são úteis quando você quer atribuir usuários a versões. A seleção aleatória é útil quando você quer distribuir a carga de trabalho uniformemente.

Ao utilizar a divisão de tráfego por endereço IP, o App Engine cria um hash — isto é, um número gerado com base em uma string de entrada entre 0 e 999, usando o endereço IP de cada versão. Isso pode criar problemas se os usuários mudarem de endereço IP, como se eles começarem a trabalhar com o aplicativo no escritório e depois mudarem para uma rede em uma cafeteria. Se a informação de estado é mantida em uma versão, ela pode não estar disponível após uma mudança de endereço IP.

A maneira preferida de dividir o tráfego é com um cookie. Quando você usa um cookie, o cabeçalho de solicitação HTTP para um cookie chamado GOOGAPPUID contém um valor hash entre 0 e 999. Com a divisão por cookie, um usuário acessará a mesma versão do aplicativo mesmo se o endereço IP do usuário mudar. Se não houver cookie GOOGAPPUID, o tráfego é roteado aleatoriamente.

O comando para dividir o tráfego é gcloud app services set-traffic. Aqui está um exemplo:

```
gcloud app services set-traffic serv1 --splits v1=.4,v2=.6
```

Este comando dividirá o tráfego, com 40% indo para a versão 1 do serviço chamado serv1 e 60% indo para a versão 2. Se nenhum nome de serviço for especificado, então todos os serviços são divididos.

O comando gcloud app services set-traffic aceita os seguintes parâmetros:

- --migrate indica que o App Engine deve migrar o tráfego da versão anterior para a nova versão.
- --split-by especifica como dividir o tráfego usando IP ou cookies. Os valores possíveis são ip, cookie e random.

Você também pode migrar o tráfego pelo console. Navegue até a página de Versões e selecione o comando Migrar.

Resumo

O Cloud Run é um serviço gerenciado e sem servidor para executar aplicativos contêinerizados. O Cloud Run suporta qualquer aplicativo que possa ser executado em um contêiner. Os serviços Cloud Run são usados quando seu código é usado para responder a solicitações da web ou eventos. Os trabalhos Cloud Run são usados quando o código executa até que uma carga de trabalho esteja completa. Ao trabalhar com serviços ou trabalhos, você pode configurar várias categorias de parâmetros, incluindo configurações de contêiner, conexão e segurança.

O App Engine Standard é uma plataforma sem servidor para executar aplicativos em ambientes específicos de linguagem. Como engenheiro de nuvem, espera-se que você saiba como implantar e escalar aplicativos App Engine. Aplicações App Engine consistem em serviços, versões e instâncias. Você pode ter várias versões em execução ao mesmo tempo. Você pode dividir o tráfego entre versões e fazer com que todo o tráfego migre automaticamente para uma nova versão. As aplicações App Engine são configuradas através de arquivos de configuração app.yaml. Você pode especificar o ambiente de linguagem, parâmetros de escalonamento e outros parâmetros para personalizar sua implantação. O App Engine não está mais listado no Guia de Exame do Google Cloud Associate Cloud Engineer, mas é incluído aqui porque espera-se que os engenheiros de nuvem estejam familiarizados com este popular serviço do Google Cloud.

Essenciais para o Exame

Seja capaz de descrever o Cloud Run como um serviço sem servidor para executar contêineres. Cloud Run é um serviço gerenciado e sem servidor para implantar, escalar e gerenciar serviços. Embora não haja servidores para configurar, você pode especificar parâmetros para controlar o número de instâncias em execução a qualquer momento, a segurança usada para proteger o serviço, bem como detalhes da configuração de conexão.

Saiba como os serviços Cloud Run são usados para executar serviços de longa duração, como sites e servidores de API. Os serviços Cloud Run executam contêineres

continuamente. Você tem a opção de pagar apenas pelos recursos de CPU usados ao responder a solicitações, ou pode optar por ter um contêiner sempre disponível e pagar pelo tempo em que os recursos de CPU estão alocados.

Saiba como os trabalhos Cloud Run são usados para executar tarefas, como carregar dados em um banco de dados. Os trabalhos Cloud Run são configurados de maneira semelhante aos serviços Cloud Run. Você pode especificar que os trabalhos usem vários contêineres executando simultaneamente. Isso é útil ao executar cargas de trabalho paralelizáveis.

Seja capaz de descrever a estrutura das aplicações App Engine Standard. Estas consistem em serviços, versões e instâncias. Os serviços geralmente fornecem uma única função. As versões são diferentes versões do código executando no ambiente App Engine. As instâncias são instâncias gerenciadas executando o serviço.

Saiba como implantar um aplicativo App Engine. Isso inclui configurar o ambiente App Engine usando o arquivo app.yaml. Saiba que um projeto pode ter apenas um aplicativo App Engine por vez. Saiba como usar o comando gcloud app deploy.

Entenda as várias opções de escalonamento. Três opções de escalonamento são escalonamento automático, escalonamento básico e escalonamento manual. Apenas o escalonamento automático e o escalonamento básico são dinâmicos. O escalonamento manual cria instâncias residentes. O escalonamento automático permite mais opções de configuração do que o escalonamento básico.

Questões

1. Você deseja fornecer aos seus clientes uma API para permitir que eles consultem um banco de dados com dados proprietários do setor. Você quer que seus desenvolvedores se concentrem em adicionar novos recursos e não na administração de servidores. Qual dos seguintes serviços do Google Cloud você escolheria?
 - A. Grupos de instâncias gerenciadas do Compute Engine
 - B. Grupos de instâncias não gerenciadas do Compute Engine
 - C. Serviços Cloud Run
 - D. Trabalhos Cloud Run
2. Você está trabalhando para um grupo de pesquisa biomédica que possui várias centenas de arquivos de dados armazenados no Cloud Storage. Eles têm um programa de análise estatística que analisa um arquivo de dados e escreve a saída para outro bucket do Cloud Storage. Eles concordaram com você que a implantação do programa em um contêiner é a melhor opção, mas estão indecisos sobre qual serviço do Google Cloud usar para executar o contêiner. O que você recomendaria?
 - A. Kubernetes Engine
 - B. Compute Engine
 - C. App Engine Flexível
 - D. Trabalhos Cloud Run
3. Você está trabalhando para um grupo de pesquisa sobre mudança climática que possui dezenas de milhares de arquivos de dados públicos de clima armazenados no Cloud Storage. Eles estão construindo um modelo para prever os níveis do mar no futuro próximo. Os dados em cada arquivo podem ser analisados independentemente dos outros arquivos. Eles planejam usar trabalhos Cloud Run para esta tarefa. Qual característica dos trabalhos Cloud Run você recomendaria que eles usassem?
 - A. Chaves de criptografia gerenciadas pelo cliente
 - B. Trabalhos em array
 - C. Conexão Cloud SQL
 - D. Um endereço IP privado
4. Um administrador de aplicativos pediu sua ajuda para configurar um serviço Cloud Run. O administrador do aplicativo gostaria de ter todas as solicitações de clientes roteadas para o mesmo contêiner, se possível. Como você sugeriria que o administrador conseguisse isso?
 - A. Usar Conexão Cloud SQL.

- B. Usar trabalhos em array.
 - C. Configurar a conexão no Serviço Cloud Run para suportar afinidade de sessão.
 - D. Usar um endereço IP privado.
5. Você está implantando um serviço no Cloud Run. O serviço tem acesso a informações pessoais identificáveis (PII) e, por motivos de conformidade, você não deseja expor o serviço a nenhum tráfego fora do tráfego interno no seu ambiente Google Cloud. Qual configuração de ingresso você usaria?
- A. Interno
 - B. Interno e Balanceamento de Carga na Nuvem
 - C. Todos
 - D. Tráfego proxy de PII
6. Você deseja usar uma conta de serviço especificamente criada para um serviço Cloud Run. Onde você especificaria isso no console da nuvem?
- A. Na aba Conexões
 - B. Na aba Segurança
 - C. Na aba Contêiner
 - D. Na aba Variáveis & Segredos
7. Um grupo de desenvolvedores precisa da capacidade de implantar novas versões de um serviço executado no Cloud Run. Como você configuraria esse acesso?
- A. Usando IAM
 - B. Usando Cloud Identity Aware Proxy (IAP)
 - C. Usando uma política de ingresso
 - D. Usando a aba Segurança no console do Cloud Run
8. Sua equipe implantou um serviço Cloud Run no mês passado que acessa um banco de dados Cloud SQL. A equipe do banco de dados mudou seu sistema e agora usa um cache Memcached executando no Cloud Memorystore. Você precisa mudar seu serviço Cloud Run para acessar o cache Cloud Memorystore. O que você usaria para fazer isso?
- A. Conexão Cloud SQL
 - B. Proxy Cloud IAP
 - C. Conexão VPC

D. Afinidade de sessão

9. Um serviço é implantado nos serviços Cloud Run e se comunicará com clientes usando gRPC. O que você deve configurar para habilitar esse protocolo a trabalhar com o serviço?

- A. Balanceamento de Carga Externo
- B. Cloud Identity Aware Proxy (IAP)
- C. Afinidade de sessão
- D. HTTP/2 de ponta a ponta

10. Quais serviços Google Cloud podem ser usados para armazenar e acessar imagens de contêiner acessíveis a partir do Cloud Run?

- A. Somente o Container Registry
- B. Container Registry e Artifact Registry
- C. Somente o Artifact Registry
- D. Container Registry, Artifact Registry e Kubernetes Engine

11. Você projetou um microsserviço que deseja implantar em produção. Antes que ele possa ser implantado, você precisa revisar como gerenciará o ciclo de vida do serviço. O arquiteto está particularmente preocupado sobre como você implantará atualizações no serviço com mínima interrupção. Que aspecto dos componentes do App Engine você usaria para minimizar interrupções durante as atualizações do serviço?

- A. Serviços
- B. Versões
- C. Grupos de instâncias
- D. Instâncias

12. Você acabou de lançar uma aplicação rodando no App Engine Standard. Você percebe que existem períodos de demanda de pico nos quais você precisa de até 12 instâncias, mas na maior parte do tempo 5 instâncias são suficientes. Qual é a melhor maneira de garantir que você tenha instâncias suficientes para atender à demanda sem gastar mais do que o necessário?

- A. Configure seu aplicativo para autoescala e especifique instâncias máximas de 12 e instâncias mínimas de 5.

- B. Configure seu aplicativo para escalonamento básico e especifique instâncias máximas de 12 e instâncias mínimas de 5.
- C. Crie um trabalho cron para adicionar instâncias pouco antes dos períodos de pico e remover instâncias depois que o período de pico terminar.
- D. Configure seu aplicativo para detecção de instância e não especifique um número máximo ou mínimo de instâncias.
13. Qual comando você deve usar para implantar um aplicativo App Engine a partir da linha de comando?
- A. gcloud components app deploy
 - B. gcloud app deploy
 - C. gcloud components instances deploy
 - D. gcloud app instance deploy
14. Você implantou um aplicativo Python Django 1.5 no App Engine. Esta versão do Django requer Python 3. Por alguma razão, o App Engine está tentando executar o aplicativo usando Python 2. Qual arquivo você verificaria e possivelmente modificaria para garantir que o Python 3 seja usado com este aplicativo?
- A. app.config
 - B. app.yaml
 - C. services.yaml
 - D. deploy.yaml
15. Você está preocupado que à medida que os usuários fazem conexões com seu aplicativo, o desempenho degradará. Você quer garantir que mais instâncias sejam adicionadas ao seu aplicativo App Engine quando houver mais de 20 solicitações simultâneas. Qual parâmetro você especificaria em app.yaml?
- A. max_concurrent_requests
 - B. target_throughput_utilization
 - C. max_instances
 - D. max_pending_latency
16. Quais parâmetros podem ser configurados com o escalonamento básico?
- A. max_instances e min_instances
 - B. idle_timeout e min_instances
 - C. idle_timeout e max_instances
 - D. idle_timeout e target_throughput_utilization
17. O parâmetro runtime em app.yaml é usado para especificar o quê?

- A. O script a executar
 - B. A URL para acessar o aplicativo
 - C. O ambiente de execução da linguagem
 - D. O tempo máximo que um aplicativo pode rodar
18. Você trabalha para uma startup e os custos são uma grande preocupação. Você está disposto a aceitar uma pequena queda de desempenho se isso lhe economizar dinheiro. Como você deve configurar o escalonamento para seus aplicativos executados no App Engine?
- A. Use instâncias dinâmicas especificando escalonamento automático ou básico.
 - B. Use instâncias residentes especificando escalonamento automático ou básico.
 - C. Use instâncias dinâmicas especificando escalonamento manual.
 - D. Use instâncias residentes especificando escalonamento manual.
19. Qual parâmetro para gcloud app services set-traffic é usado para especificar o método a ser usado ao dividir o tráfego?
- A. —split-traffic
 - B. —split-by
 - C. —traffic-split
 - D. —split-method
20. Quais são os métodos válidos para dividir o tráfego no App Engine?
- A. Apenas por endereço IP
 - B. Apenas por cookie HTTP
 - C. Aleatoriamente e apenas por endereço IP
 - D. Por endereço IP, cookies HTTP e aleatoriamente

Capítulo 10

Computação com Funções de Nuvem

ESTE CAPÍTULO COBRE OS SEGUINTESS OBJETIVOS DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 3.3 Implantação e implementação de recursos do Cloud Run e Cloud Functions

Neste capítulo, descrevemos o propósito das Cloud Functions, bem como como implementar e implantar as funções. Usaremos exemplos de funções escritas em Python. Se você não está familiarizado com Python, isso não deve desencorajá-lo a acompanhar, pois explicaremos os detalhes importantes das funções Python. Você aprenderá a usar o Cloud Console e comandos gcloud para criar e gerenciar Cloud Functions.

Introdução às Cloud Functions

Cloud Functions é um serviço de computação sem servidor fornecido pelo Google Cloud. Cloud Functions é semelhante ao Cloud Run, pois ambos são opções de computação sem servidor. Uma diferença primária é que o Cloud Run suporta tanto serviços com endpoints HTTP que podem rodar continuamente quanto trabalhos em lote que rodam até a conclusão e terminam, enquanto as Cloud Functions são funções de execução relativamente curta (até 60 minutos para funções HTTP e 10 minutos para funções acionadas por eventos).

Cloud Functions são bem adequadas para processamento baseado em eventos. Por exemplo, seus clientes podem carregar arquivos para o Cloud Storage, que são analisados para verificações de controle de qualidade, e se as verificações forem aprovadas, uma mensagem é escrita em um tópico do Pub/Sub, um serviço de mensagens no GCP, que é lido por outro serviço que continua o processamento.

No momento da escrita, existem duas versões suportadas de Cloud Functions: as Cloud Functions originais e as Cloud Functions de Segunda Geração. As Cloud Functions de Segunda Geração oferecem instâncias maiores, melhor concorrência, instâncias pré-aquecidas e gestão de tráfego. As Funções de Segunda Geração também suportam o Eventarc, um serviço do GCP que suporta a gestão do fluxo de eventos em arquiteturas de microsserviços. O Eventarc expande grandemente a gama de fontes de eventos suportadas pelas Cloud Functions. As funções de Segunda Geração também suportam CloudEvents, uma especificação aberta para descrever eventos em nuvem.

Eventos, Gatilhos e Funções

Aqui estão alguns termos que você precisa conhecer antes de prosseguir com Cloud Functions:

- Eventos
- Gatilhos
- Funções

Eventos são uma ação particular que acontece na nuvem, como um arquivo sendo carregado para o Cloud Storage ou uma mensagem que é escrita em uma fila de mensagens do Pub/Sub (chamada de tópico). Existem diferentes tipos de ações associadas a cada um dos eventos. A primeira geração de Cloud Functions, GCP suporta eventos em várias categorias:

- HTTP
- Cloud Storage
- Cloud Pub/Sub
- Cloud Firestore
- Cloud Firebase

O tipo de evento HTTP permite que desenvolvedores invoquem uma função fazendo uma solicitação HTTP usando chamadas POST, GET, PUT, DELETE e OPTIONS. Eventos no Cloud Storage incluem carregar, deletar e arquivar um arquivo. Cloud Pub/Sub tem um evento para publicar uma mensagem. Cloud Firestore é um banco de dados de documentos NoSQL, e Cloud Functions suporta eventos de criar, atualizar, deletar e escrever. Firebase é um serviço de banco de dados usado para o desenvolvimento de aplicativos móveis e suporta gatilhos de banco de dados, gatilhos de configuração remota e gatilhos de autenticação.

Segunda Geração Cloud Functions usa gatilhos do Eventarc, que são configurados com base em um provedor, como serviços suportados nas Cloud Functions de Primeira Geração como Cloud Pub/Sub; serviços adicionais do GCP, como Cloud Task, Cloud Dataproc, Cloud DNS e Gerenciamento de Rede; bem como serviços específicos não GCP como OAuth 2.0. Os tipos específicos de eventos variam de acordo com o provedor. Por exemplo, provedores de OAuth 2.0 suportam eventos GetToken, GetTokenInfo e RevokeToken. Eventos de Gerenciamento de Rede incluem CreateConnectivityTest, GetConnectivityTest, ListConnectivityTest.

Para cada um dos eventos habilitados para Cloud Functions que podem ocorrer, você pode definir um gatilho. Um gatilho é uma maneira de responder a um evento. Gatilhos têm uma função associada. A função, passada argumentos com dados sobre o evento, executa em resposta ao evento.

Ambientes de Execução

As funções são executadas em seu próprio ambiente. Cada vez que uma função é invocada, ela é executada em uma instância separada de todas as outras invocações. Não há como compartilhar informações entre invocações de funções usando apenas Cloud Functions. Se você precisa coordenar a atualização de dados, como manter uma contagem global, ou precisa manter informações sobre o estado das funções, como o nome do último evento processado, então você deve usar um banco de dados como o Cloud Firestore ou um arquivo no Cloud Storage.

O Google atualmente suporta vários ambientes de execução:

- Node.js
- Python
- Go
- Java
- .NET
- Ruby
- PHP

Para cada um desses ambientes de execução, versões específicas podem ser recomendadas em detrimento de outras. Por exemplo, no momento da escrita, o Node.js recomendado é o Node.js 16 e a versão recomendada do Python é 3.9. Consulte a

documentação das Cloud Functions (<https://cloud.google.com/functions>) para as últimas versões suportadas, recomendadas e obsoletas dos ambientes de execução.

Vamos passar por um exemplo de função. Digamos que você queira registrar informações sobre uploads de arquivos para um determinado bucket no Cloud Storage. Isso pode ser feito escrevendo uma função Python que recebe informações sobre um evento e emite comandos de impressão para enviar uma descrição desses dados para um arquivo de log. Aqui está o código Python:

```
def cloud_storage_function_test(event_data, event_context):  
    print('ID do Evento: {}'.format(event_context.event_id))  
    print('Tipo de Evento: {}'.format(event_context.event_type))  
    print('Arquivo: {}'.format(event_data['name']))
```

A primeira linha inicia a criação de uma função chamada `cloud_storage_function_test`. Ela recebe dois argumentos, `event_data` e `event_context`. Estas são estruturas de dados Python com informações sobre o objeto do evento e sobre o próprio evento. As próximas três linhas imprimem os valores do `event_id`, `event_type` e nome do arquivo. Como este código será executado como uma função e não interativamente, a saída de um comando `print` irá para o arquivo de log da função.

Funções Python devem ser salvas em um arquivo chamado `main.py`.

Mundo Real

Tornando Documentos Pesquisáveis

Litígios, ou processos judiciais, entre empresas muitas vezes envolvem a revisão de um grande volume de documentos. Documentos eletrônicos podem estar em formatos facilmente pesquisáveis, como documentos do Microsoft Word ou arquivos PDF. Outros podem ser imagens digitalizadas de documentos em papel. Nesse caso, o arquivo precisa ser pré-processado usando um programa de reconhecimento óptico de caracteres (OCR). Funções podem ser usadas para automatizar o processo de OCR. Quando um arquivo é carregado, um gatilho do Cloud Storage é acionado e invoca uma função. A função determina se o arquivo está em um formato pesquisável ou precisa ser pré-processado pelo programa OCR. Se o arquivo exigir processamento OCR, a função escreve a localização do arquivo em um tópico do Pub/Sub.

Uma segunda função é vinculada a um novo evento de mensagem. Quando a localização de um arquivo é escrita em uma mensagem, a função chama o programa OCR para escanear o documento e produzir uma versão pesquisável do arquivo. Essa versão pesquisável é escrita em um bucket do Cloud Storage, onde pode ser indexada pela ferramenta de pesquisa junto com outros arquivos pesquisáveis.

Funções de Nuvem Recebendo Eventos do Cloud Storage

O Cloud Storage é o serviço de armazenamento de objetos do GCP. Este serviço permite que você armazene arquivos em contêineres conhecidos como buckets. Iremos detalhar mais sobre o Cloud Storage no Capítulo 11, "Planejando o Armazenamento na

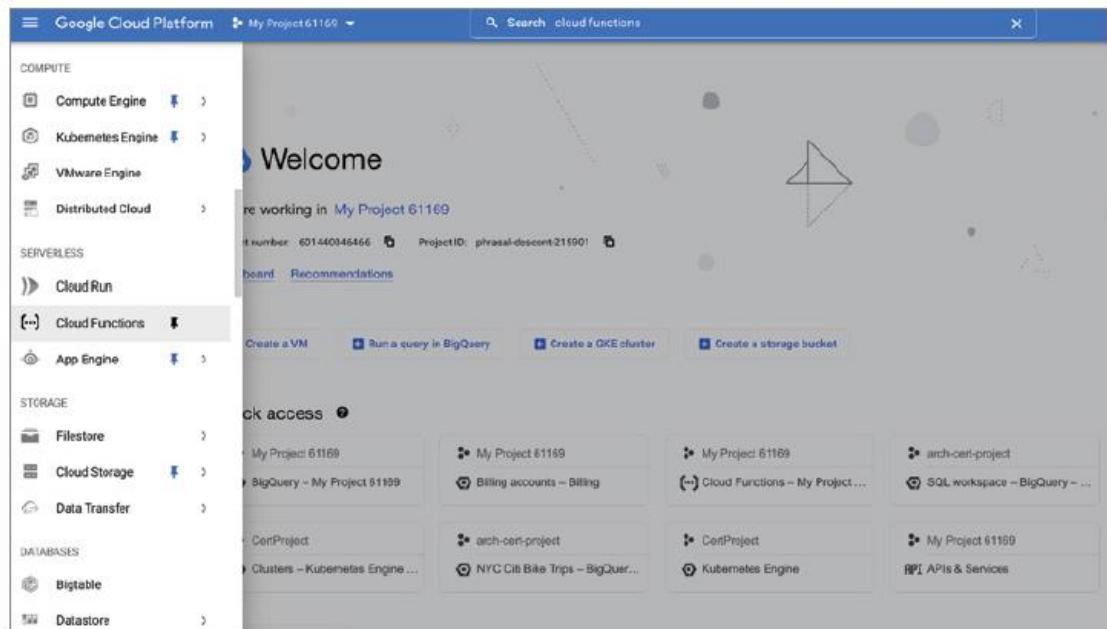
Nuvem", mas para este capítulo, você só precisa entender que o Cloud Storage usa buckets para armazenar arquivos. Quando arquivos são criados, deletados, arquivados ou suas metadatas mudam, um evento pode invocar uma função. Vamos passar por um exemplo de implantação de uma função para Eventos do Cloud Storage usando o Cloud Console e comandos gcloud no Cloud SDK e Cloud Shell.

Implantando uma Função de Nuvem para Eventos do Cloud Storage

Usando o Cloud Console

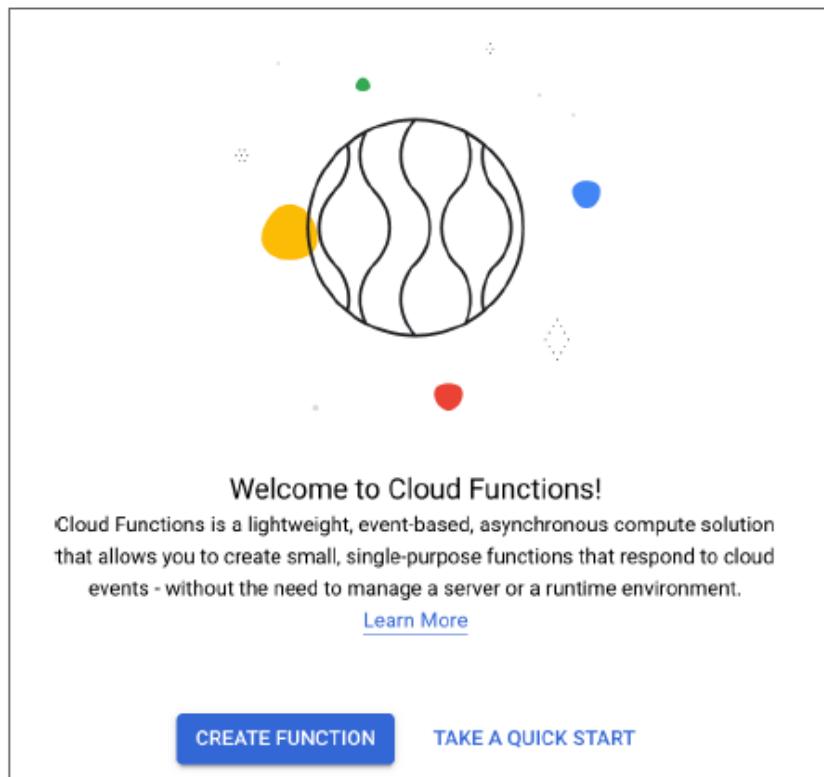
Para criar uma função usando o Cloud Console, selecione as opções de Função de Nuvem no menu vertical do console, conforme mostrado na Figura 10.1.

FIGURE 10.1 Opening the Cloud Functions console



No console de Funções de Nuvem, pode ser que você seja solicitado a ativar a API de Funções de Nuvem se ela ainda não estiver ativada. Após a API ser ativada, você terá a opção de criar uma nova função, conforme mostrado na Figura 10.2.

FIGURE 10.2 The Create Function button in Cloud Console



Quando você cria uma nova função no console, um formulário como o da Figura 10.3 aparece. Na Figura 10.3, as opções, que foram preenchidas, incluem:

- Nome da função
- Região
- Tipo de gatilho
- Tipo de evento
- Bucket

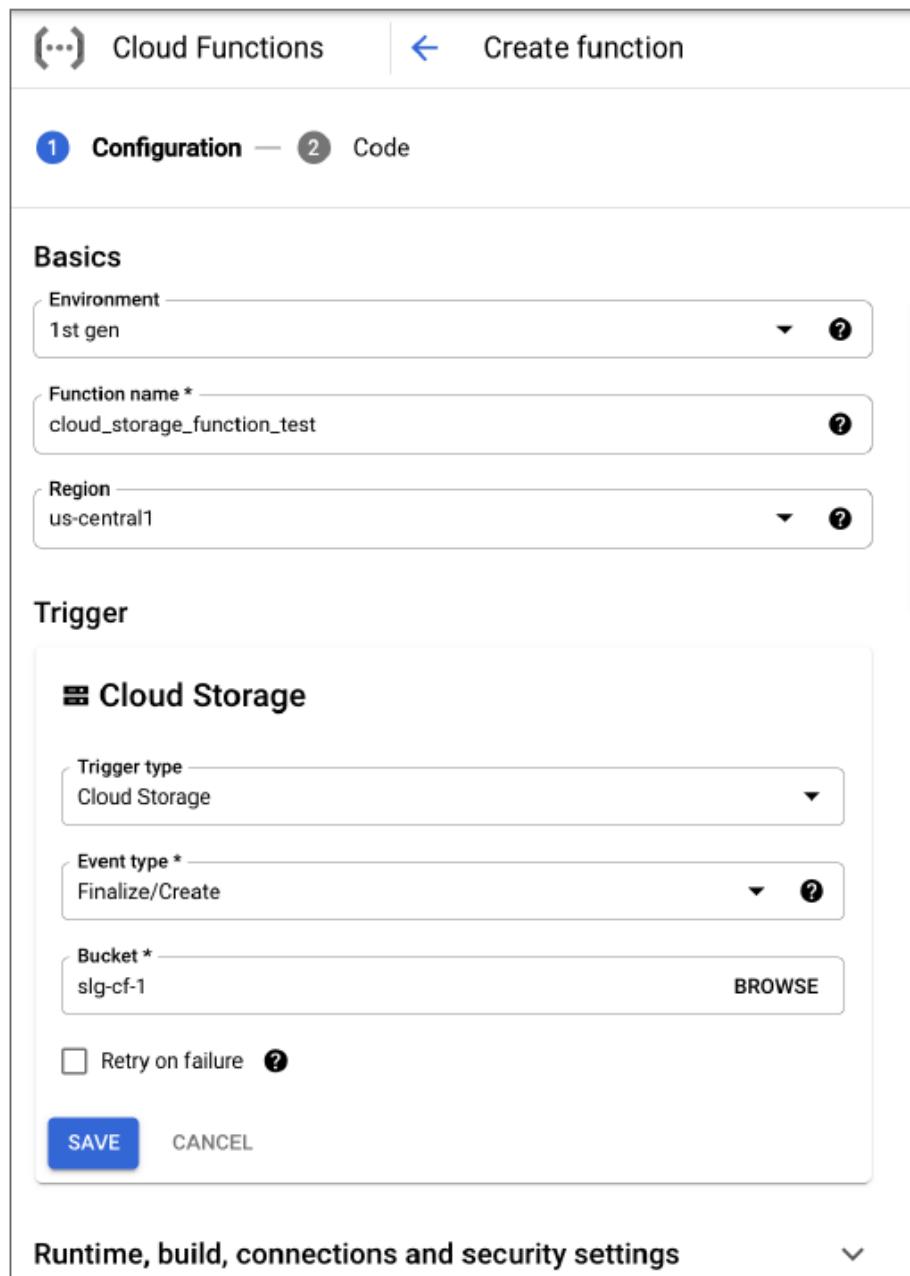
No exemplo a seguir, estamos fazendo o upload de um arquivo contendo o código da função. O conteúdo desse arquivo é o seguinte:

```
def cloud_storage_function_test(event_data, event_context):  
    print('ID do Evento: {}'.format(event_context.event_id))  
    print('Tipo de Evento: {}'.format(event_context.event_type))  
    print('Arquivo: {}'.format(event_data['name']))
```

O nome da função é como o Google Cloud se referirá a esta função. Memória Alocada é a quantidade de memória que estará disponível para a função. As opções de memória variam de 128 MB a 8 GB para as Funções de Nuvem originais e 16 GB com as Funções de Nuvem de Segunda Geração. Gatilho é um dos gatilhos definidos, como HTTP, Cloud Pub/Sub e Cloud Storage. Há várias opções para especificar onde encontrar

o código-fonte, incluindo fazer o upload dele, obtê-lo do Cloud Storage ou de um repositório Cloud Source, ou inserir o código em um editor. Runtime indica qual ambiente de execução usar para executar o código. O editor é onde você pode inserir o código da função. Finalmente, a função a executar é o nome da função no código que deve ser executada quando o evento ocorrer.

FIGURE 10.3 Creating a function in the console



Após criar uma função, você verá uma lista de funções no console do Cloud Functions, conforme mostrado na Figura 10.4.

FIGURE 10.4 List of functions in the console

Environment	Name	Last deployed	Region	Trigger	Runtime	Memory allocated
1st gen	function-1	Nov 19, 2022, 9:54:25 AM	us-central1	HTTP	Node.js 16	256 MB

Implantando uma Função de Nuvem para Eventos do Cloud Storage

Usando Comandos gcloud

O primeiro passo para usar comandos gcloud para Funções de Nuvem é garantir que você tenha a versão mais recente dos comandos instalada. Você pode atualizar os comandos gcloud padrão usando isso:

`gcloud components update`

Se algum dos comandos para o ambiente que você escolheu estiver em beta, você pode garantir que eles estejam instalados com o seguinte comando:

`gcloud components install beta`

Vamos supor que você tenha criado um bucket do Cloud Storage chamado `gcp-ace-exam-test-bucket`. Você pode implantar uma função usando o comando `gcloud functions deploy`. Este comando leva o nome de uma função como seu argumento. Há também três parâmetros que você precisará passar:

- runtime
- trigger-resource
- trigger-event

`runtime` indica se você está usando Python 3.7, Node.js 6 ou Node.js 8. `trigger-resource` indica o nome do bucket associado ao gatilho. `trigger-event` é o tipo de evento que acionará a execução da função. As opções possíveis são as seguintes:

- `google.storage.object.finalize`
- `google.storage.object.delete`
- `google.storage.object.archive`
- `google.storage.object.metadataUpdate`

`finalize` é o termo usado para descrever quando um arquivo é totalmente carregado.

Sempre que um novo arquivo for carregado no bucket chamado gcp-ace-exam-test-bucket, queremos executar o cloud_storage_function_test. Isso é realizado emitindo o seguinte comando:

```
gcloud functions deploy cloud_storage_function_test \
--runtime python39 \
--trigger-resource gcp-ace-exam-test-bucket \
--trigger-event google.storage.object.finalize
```

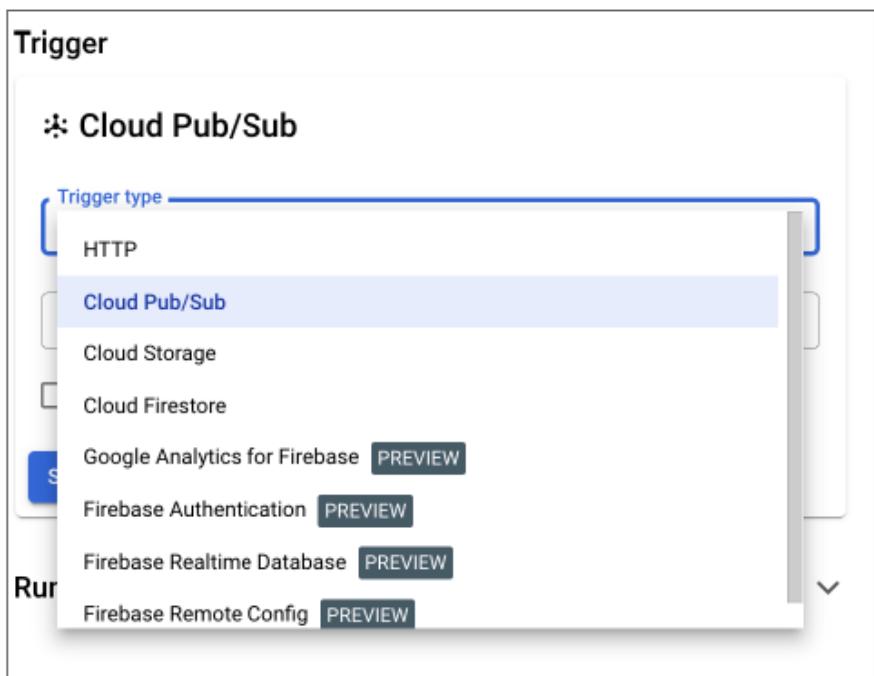
Quando você carrega um arquivo no bucket, a função executará e criará uma mensagem de log. Quando você terminar com a função e quiser deletá-la, você pode usar o comando de delete do gcloud functions, assim:

```
gcloud functions delete cloud_storage_function_test
```

Essa função imprime o ID do evento e o tipo de evento associados à mensagem. Se os dados do evento tiverem um par chave-valor com a chave de nome, então a função também imprimirá o nome na mensagem. Note que essa função tem uma declaração de importação e usa uma função chamada base64.b64decode. Isso acontece porque as mensagens no Pub/Sub são codificadas para permitir dados binários em um local onde se espera dados de texto, e a função base64.b64decode é usada para convertê-la em uma codificação de texto mais comum chamada UTF-8.

O código é implantado da mesma maneira que o exemplo anterior do Cloud Storage com duas exceções. Em vez de selecionar um gatilho do Cloud Storage, escolha Cloud Pub/Sub na lista de gatilhos, conforme mostrado na Figura 10.5.

FIGURE 10.5 Selecting a trigger from options in Cloud Console



Você também pode especificar o nome do tópico do Cloud Pub/Sub após especificar que este é um gatilho do Cloud Pub/Sub. Se o tópico não existir, ele pode ser criado, conforme mostrado na Figura 10.6.

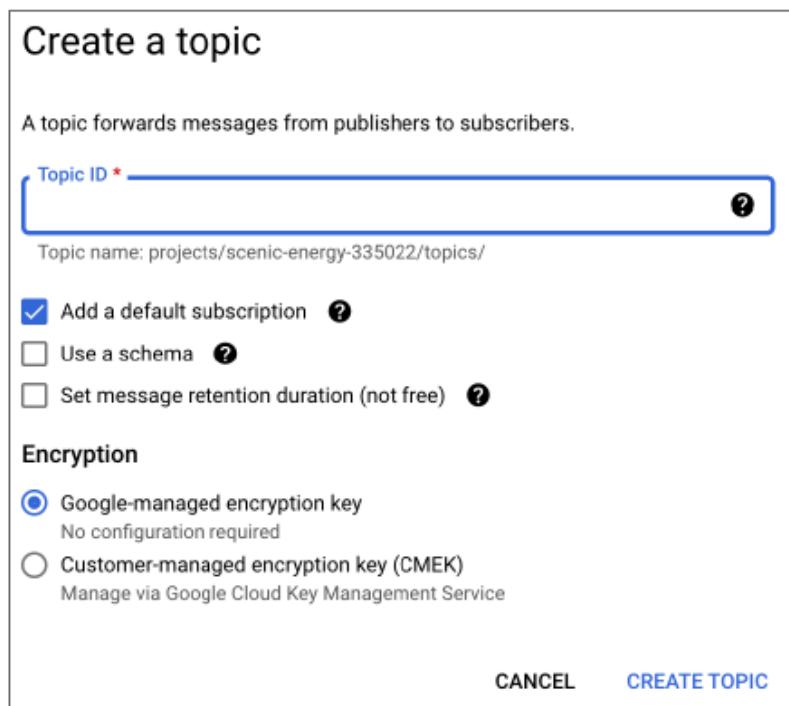
Implantando uma Função de Nuvem para Eventos do Cloud Pub/Sub

Usando Comandos gcloud

Para implantar esta função, você usa o comando `gcloud functions deploy`. Ao implantar uma função do Cloud Pub/Sub, você especifica o nome do tópico que conterá mensagens que acionarão a função. Como na implantação para o Cloud Storage, você tem que especificar o ambiente de execução que deseja usar. Aqui está um exemplo:

```
gcloud functions deploy pub_sub_function_test --runtime python39 --trigger-topic gcp-ace-exam-test-topic
```

FIGURE 10.6 Creating a Pub/Sub topic while creating a Cloud Function



Você pode deletar essa função usando o comando `gcloud functions delete`. Aqui está um exemplo:

```
gcloud functions delete pub_sub_function_test
```

Resumo

Neste capítulo, trabalhamos com Cloud Functions e vimos como implementar e implantar funções. Usamos exemplos de funções escritas em Python, mas elas poderiam ter sido escritas em Node.js ou em um dos vários outros idiomas suportados também. Funções podem ser criadas usando o Google Cloud Console ou a linha de comando. Para usar Cloud Functions, é importante entender a relação entre eventos, gatilhos e funções.

Eventos são ações que acontecem na nuvem. Diferentes serviços têm diferentes tipos de eventos. Gatilhos são como você indica que deseja executar uma função quando um evento ocorre. Funções referem-se ao código que é executado quando um evento ocorre e tem um gatilho definido para ele.

Essenciais para o Exame

Conheça a relação entre eventos, gatilhos e funções. Eventos são ações que acontecem, como quando um arquivo é carregado no Cloud Storage ou uma mensagem é escrita em um tópico do Cloud Pub/Sub. Gatilhos são declarações de que uma ação deve ser tomada quando um evento ocorre. Funções associadas a gatilhos definem quais ações são tomadas quando um evento ocorre.

Saiba quando usar Cloud Functions. Cloud Functions é um serviço que suporta funções de propósito único que respondem a eventos na nuvem. Cloud Run também é uma opção de computação sem servidor, mas é usada para implantar aplicações multifuncionais, incluindo aquelas com as quais os usuários interagem diretamente.

Conheça os ambientes de execução e gerações suportadas em Cloud Functions. Cloud Functions suporta os seguintes ambientes de execução: Node.js, Python, Go, Java, .NET, Ruby e PHP. Existem duas gerações de Cloud Functions, a original é conhecida como Primeira Geração e a outra Segunda Geração. Cloud Functions de Segunda Geração têm menos restrições e mais funcionalidades.

Conheça os parâmetros para definir uma função de nuvem em um evento do Cloud Storage. Os parâmetros para Cloud Storage incluem os seguintes:

- Nome da função de nuvem
- Memória alocada para a função
- Gatilho
- Tipo de evento
- Fonte do código da função
- Ambiente de execução
- Código-fonte
- Nome da função a executar

Conheça os parâmetros para definir uma Cloud Function em um evento do Cloud Pub/Sub. Os parâmetros para Pub/Sub incluem os seguintes:

- Nome da função de nuvem
- Memória alocada para a função
- Gatilho
- Tópico

- Fonte do código da função
- Ambiente de execução
- Código-fonte
- Nome da função a executar

Conheça os comandos `gcloud` para trabalhar com Cloud Functions. Estes incluem os seguintes:

- `gcloud functions deploy`
- `gcloud functions delete`

Questões

1. Um gerente de produto está propondo um novo aplicativo que exigirá vários serviços de back-end e três serviços de lógica de negócios. Cada serviço fornecerá uma única função, e vários desses serviços serão necessários para completar uma tarefa de negócios. O tempo de execução do serviço depende do tamanho da entrada e espera-se que leve até 90 minutos em alguns casos. Qual produto do GCP é uma boa opção sem servidor para executar esse serviço relacionado?
 - A. Cloud Functions
 - B. Compute Engine
 - C. Cloud Run
 - D. Cloud Storage
2. Foi solicitado que você implantasse uma função de nuvem para reformatar arquivos de imagem assim que fossem carregados no Cloud Storage. Você percebe, após algumas horas, que cerca de 10% dos arquivos não são processados corretamente. Após revisar os arquivos que falharam, você percebe que todos são substancialmente maiores que a média. Qual poderia ser a causa das falhas?
 - A. Há um erro de sintaxe no código da função.
 - B. O runtime selecionado está errado.
 - C. O tempo de espera é muito curto para permitir tempo suficiente para processar arquivos grandes.
 - D. Há um erro de permissões no bucket do Cloud Storage que contém os arquivos.
3. Quando uma ação ocorre no GCP, como um arquivo sendo escrito no Cloud Storage ou uma mensagem sendo adicionada a um tópico do Cloud Pub/Sub, como essa ação é chamada?
 - A. Um incidente
 - B. Um evento
 - C. Um gatilho
 - D. Uma entrada de log
4. Todos os seguintes geram eventos que podem ser acionados usando Cloud Functions, exceto qual deles?
 - A. Cloud Storage
 - B. Cloud Pub/Sub
 - C. SSL
 - D. Firebase

5. Quais runtimes são suportados no Cloud Functions?
 - A. Apenas Node.js e Python
 - B. Apenas Node.js, Python e Ruby
 - C. Apenas Node.js, Python, .NET e Go
 - D. Apenas Node.js, Python, Go, Java, .NET, Ruby e PHP
6. Um gatilho HTTP pode ser invocado fazendo uma solicitação usando quais dos seguintes?
 - A. Apenas GET
 - B. POST e GET apenas
 - C. DELETE, POST e GET
 - D. DELETE, POST, REVISE e GET
7. Quais tipos de eventos estão disponíveis para Cloud Functions trabalhando com Cloud Storage?
 - A. Apenas upload ou finalização e delete
 - B. Apenas upload ou finalização, delete e listagem
 - C. Apenas upload ou finalização, delete e atualização de metadados
 - D. Upload ou finalização, delete, arquivo e atualização de metadados
8. Você tem a tarefa de projetar uma função para executar em Cloud Functions. A função precisará de mais do que a quantidade padrão de memória e deve ser aplicada apenas quando um evento de finalização ocorrer após um arquivo ser carregado no Cloud Storage. A função deve aplicar sua lógica apenas a arquivos com um tipo de arquivo de imagem padrão. Qual das seguintes características obrigatórias não pode ser especificada em um parâmetro e deve ser implementada no código da função?
 - A. Nome da função de nuvem
 - B. Memória alocada para a função
 - C. Tipo de arquivo para aplicar a função
 - D. Tipo de evento
9. Quanta memória pode ser alocada para uma Cloud Function ao usar funções de Segunda Geração?
 - A. 128 MB a 256 MB
 - B. 128 MB a 512 MB
 - C. 128 MB a 1 GB
 - D. 128 MB a 16 GB

10. Por quanto tempo uma função de Cloud Function de tipo de evento de Segunda Geração pode ser executada por padrão antes de expirar?
- A. 30 segundos
 - B. 1 minuto
 - C. 10 minutos
 - D. 20 minutos
11. Você deseja usar a linha de comando para gerenciar Cloud Functions que serão escritas em Python. Qual comando você deve executar para garantir que seu SDK de linha de comando esteja atualizado?
- A. gcloud components install
 - B. gcloud install components functions
 - C. gcloud functions install components
 - D. gcloud functions install
12. Você deseja criar uma função na nuvem para transformar arquivos de áudio em diferentes formatos. Os arquivos de áudio serão carregados no Cloud Storage. Você quer iniciar as transformações assim que os arquivos terminarem de ser carregados. Qual gatilho você especificaria na Cloud Function para fazê-la executar após o arquivo ser carregado?
- A. google.storage.object.finalize
 - B. google.storage.object.upload
 - C. google.storage.object.archive
 - D. google.storage.object.metadataUpdate
13. Você está definindo uma Cloud Function para escrever um registro em um banco de dados quando um arquivo no Cloud Storage é arquivado. Quais parâmetros você terá que definir ao criar essa função?
- A. apenas runtime
 - B. apenas trigger-resource
 - C. apenas runtime, trigger-resource, trigger-event
 - D. runtime, trigger-resource, trigger-event, file-type
14. Você gostaria de parar de usar uma Cloud Function e deletá-la do seu projeto. Qual comando você usaria da linha de comando para deletar uma Cloud Function?
- A. gcloud functions delete
 - B. gcloud components function delete

- C. gcloud components delete
 - D. gcloud delete functions
15. Pediram para você implantar uma Cloud Function para trabalhar com Cloud Pub/Sub. Ao revisar o código Python, você nota uma referência a uma função Python chamada base64.b64decode. Por que uma função de decodificação seria necessária em uma cloud function do Pub/Sub?
- A. Não é necessário e não deveria estar lá.
 - B. As mensagens em tópicos do Pub/Sub são codificadas para permitir que dados binários sejam usados onde se espera dados de texto. As mensagens precisam ser decodificadas para acessar os dados na mensagem.
 - C. É necessário adicionar caracteres de preenchimento ao final da mensagem para tornar todas as mensagens do mesmo comprimento.
 - D. A função de decodificação mapeia dados de uma estrutura de dados de dicionário para uma estrutura de dados de lista.
16. Qual destes comandos implantará uma Cloud Function Python chamada pub_sub_function_test?
- A. gcloud functions deploy pub_sub_function_test
 - B. gcloud functions deploy pub_sub_function_test --runtime python37
 - C. gcloud functions deploy pub_sub_function_test --runtime python37 --trigger-topic gcp-ace-exam-test-topic
 - D. gcloud functions deploy pub_sub_function_test --runtime python --trigger-topic gcp-ace-exam-test-topic
17. Ao especificar uma Cloud Function do Cloud Storage, você deve especificar um tipo de evento, como finalizar, deletar ou arquivar. Ao especificar uma Cloud Function do Cloud Pub/Sub, você não precisa especificar um tipo de evento. Por que isso acontece?
- A. Cloud Pub/Sub não possui gatilhos para tipos de evento.
 - B. Cloud Pub/Sub tem gatilhos apenas para um tipo de evento, quando uma mensagem é publicada.
 - C. Cloud Pub/Sub determina o tipo de evento correto analisando o código da função.
 - D. A afirmação na pergunta está incorreta; você precisa especificar um tipo de evento com funções do Cloud Pub/Sub.
18. Sua empresa tem um aplicativo web que permite aos candidatos a emprego fazerem upload de arquivos de currículo. Alguns arquivos estão em Microsoft Word, alguns são PDFs e outros são arquivos de texto. Você gostaria de armazenar todos os currículos como PDFs. Como você faria isso de uma maneira que

minimize o tempo entre o upload e a conversão e com quantidades mínimas de codificação?

- A. Escreva um aplicativo Cloud Run com múltiplos serviços para converter todos os documentos para PDF.
 - B. Implemente uma Cloud Function no Cloud Storage para executar em um evento de finalização. A função verifica o tipo de arquivo e, se não for PDF, a função chama uma função conversora para PDF e escreve a versão em PDF no bucket que tem o original.
 - C. Adicione os nomes de todos os arquivos a um tópico do Cloud Pub/Sub e tenha um trabalho em lote executado em intervalos regulares para converter os arquivos originais para PDF.
 - D. Implemente uma Cloud Function no Cloud Pub/Sub para executar em um evento de finalização. A função verifica o tipo de arquivo e, se não for PDF, a função chama uma função conversora para PDF e escreve a versão em PDF no bucket que tem o original.
19. Quais são as opções para fazer upload de código para uma cloud function?
- A. Editor inline
 - B. Upload de zip
 - C. Repositório de código fonte do Cloud
 - D. Todas as opções acima
20. Que tipo de gatilho permite que desenvolvedores usem chamadas HTTP POST, GET e PUT para invocar uma Cloud Function?
- A. HTTP
 - B. Webhook
 - C. Cloud HTTP
 - D. Nenhuma das opções acima

Capítulo 11

Planejando o Armazenamento na Nuvem

ESTE CAPÍTULO COBRE O SEGUINTE OBJETIVO DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 2.3 Planejamento e configuração de opções de armazenamento de dados

Como engenheiro de nuvem, você terá que entender as várias opções de armazenamento fornecidas na Plataforma Google Cloud (Google Cloud). Espera-se que você escolha a opção apropriada para um determinado caso de uso, conhecendo os compromissos relativos, como ter acesso ao SQL para uma linguagem de consulta versus a capacidade de armazenar e consultar petabytes de dados fluindo para o seu banco de dados.

Diferentemente da maioria dos outros capítulos do livro, este capítulo foca mais em conceitos de armazenamento do que em realizar tarefas específicas no Google Cloud. O material aqui ajudará você a responder perguntas sobre a escolha da melhor solução de armazenamento. O Capítulo 12, "Implantando Armazenamento no Google Cloud," fornecerá detalhes sobre a implantação e implementação de soluções de dados.

Para escolher entre as opções de armazenamento, ajuda entender como as soluções de armazenamento variam por:

- Tempo para acessar dados
- Modelo de dados
- Outras características, como consistência, disponibilidade e suporte para transações

Este capítulo inclui diretrizes para escolher soluções de armazenamento para diferentes tipos de requisitos.

Tipos de Sistemas de Armazenamento

Uma consideração principal quando você escolhe uma solução de armazenamento é o tempo em que os dados devem ser acessados. Em um extremo, dados em um cache L1 em um chip de CPU podem ser acessados em 0,5 nanosegundos (ns). No outro extremo do espectro, alguns serviços podem exigir horas para retornar arquivos de dados. A maioria dos requisitos de armazenamento fica entre esses extremos.

Nanosegundos, Milissegundos e Microsssegundos

Alguns sistemas de armazenamento operam em velocidades tão desconhecidas para nós quanto o que acontece sob um microscópio eletrônico. Um segundo é um tempo extremamente longo quando se fala sobre o tempo que leva para acessar dados na memória ou em disco. Medimos o tempo de acesso, ou "latência", com três unidades de medida:

- Nanossegundo (ns), que é 10^{-9} segundo
- Microsssegundo (μ s), que é 10^{-6} segundo
- Milissegundo (ms), que é 10^{-3} segundo

Note que o número 10^{-3} está em notação científica e significa 0,001 segundo. Similarmente, 10^{-6} é o mesmo que 0,000001, e 10^{-9} é o mesmo que 0,000000001 segundo.

Outra consideração é a persistência. Quão duráveis são os dados armazenados em um sistema específico? Caches oferecem a menor latência para acessar dados, mas esse tipo de dado volátil existe apenas enquanto energia é fornecida à memória. Desligue o servidor e seus dados se vão. Discos rígidos têm taxas de durabilidade mais altas, mas podem falhar. Redundância ajuda aqui. Fazendo cópias de dados e armazenando-os em diferentes servidores, em diferentes racks, em diferentes zonas e em diferentes regiões, você reduz o risco de perder dados devido a falhas de hardware.

O Google Cloud tem vários serviços de armazenamento, incluindo os seguintes:

- Um serviço gerenciado para caching baseado em Redis e Memcached
- Armazenamento de disco persistente para uso com VMs
- Armazenamento de objeto para acesso compartilhado a arquivos entre recursos
- Armazenamento arquivístico para requisitos de acesso a longo prazo e infrequente

Cache

Um cache é uma loja de dados em memória projetada para fornecer a aplicações acesso a dados em submilissegundos. Sua principal vantagem sobre outros sistemas de armazenamento é sua baixa latência. Caches são limitados em tamanho pela quantidade de memória disponível, e se a máquina que hospeda o cache for desligada, então o conteúdo do cache é perdido. Essas são limitações significativas, mas em alguns casos de uso, os benefícios do acesso rápido aos dados superam as desvantagens.

Memorystore

O Google Cloud oferece o Memorystore, um serviço gerenciado que fornece caching compatível com Redis ou Memcached. Tanto o Redis quanto o Memcached são sistemas de cache de código aberto amplamente utilizados. Como o Memorystore é compatível com os protocolos do Redis e Memcached, ferramentas e aplicações escritas para trabalhar com qualquer um deles devem funcionar com o Memorystore.

Caches geralmente são usados com uma aplicação que não pode tolerar longas latências ao recuperar dados. Por exemplo, uma aplicação que lê de um disco rígido pode ter que esperar 80 vezes mais do que se os dados fossem lidos de um cache em memória. Desenvolvedores de aplicativos podem usar caches para armazenar dados que são recuperados de um banco de dados e então recuperados do cache em vez do disco na próxima vez que os dados forem necessários.

Quando você usa o Memorystore, você cria instâncias que executam o Redis ou Memcached. Uma instância do Redis é configurada com até 300 GB de memória. Ela também pode ser configurada para alta disponibilidade, caso em que o Memorystore cria réplicas de failover. Instâncias do Memcached são configuradas como um conjunto de até 20 nós, e cada nó pode ter um máximo de 256 GB. Uma instância pode suportar até 5 TB de memória.

Configurando Memorystore

Caches do Memorystore podem ser usados com aplicações rodando no Compute Engine, App Engine e Kubernetes Engine. A Figura 11.1 mostra os parâmetros usados para configurar o Memorystore. Você pode navegar até este formulário escolhendo Memorystore no menu principal do console e, em seguida, selecionando a opção para criar uma instância do Redis.

FIGURE 11.1 Configuration parameters for a Memorystore Redis cache

The screenshot shows the 'Create a Redis instance' interface. It includes sections for 'Name your instance', 'Tier Selection', 'Capacity', and a 'Summary' sidebar.

Name your instance:
Instance ID: (Required)
Display name:

Tier Selection:
Determines availability, cost, and performance.
 Basic: Lower cost. Does not provide high availability.
 Standard: Supports automatic failover for high availability and up to 5 read replicas for scaling reads. [Learn more](#).

Capacity:
16 GB [Change](#)

Summary:

Tier	Standard
Location	us-central1
Estimated maximum throughput (MB/s)	1250 / 2000

Cost estimate:
Based on instance tier, region, and capacity.
[Pricing details](#)

16 GB with 2 read replicas	\$805.92/month
----------------------------	----------------

Para configurar um cache do Redis no Memorystore, você precisará especificar um ID de instância, um nome de exibição e uma versão do Redis. Você pode escolher ter uma réplica em uma zona diferente para alta disponibilidade selecionando o nível de instância Standard. O nível de instância Basic não inclui uma réplica, mas custa menos. A configuração de um Memcached é similar, mas também tem parâmetros para configurar um cluster de nós.

Você precisará especificar uma região e zona, juntamente com a quantidade de memória que deseja dedicar ao seu cache. O cache pode ter de 1 GB a 300 GB de tamanho. A instância do Redis será acessível a partir da rede padrão, a menos que você especifique uma rede diferente. (Veja o Capítulo 14, "Redes na Nuvem: Redes Privadas Virtuais e Redes Privadas Virtuais," e o Capítulo 15, "Redes na Nuvem: DNS, Balanceamento de Carga, Acesso Privado do Google e Endereçamento IP," para mais informações sobre redes no Google Cloud.) As opções avançadas para Memorystore permitem atribuir rótulos e definir um intervalo de IP do qual o endereço IP será atribuído.

A configuração de um Memcached é semelhante.

Armazenamento Persistente

No Google Cloud, os discos persistentes fornecem armazenamento de bloco durável. Discos persistentes podem ser anexados a VMs no Google Compute Engine (GCE) e no Google Kubernetes Engine (GKE). Como os discos persistentes são dispositivos de armazenamento de bloco, você pode criar sistemas de arquivos nesses dispositivos. Discos persistentes não são diretamente anexados a servidores físicos que hospedam suas VMs, mas são acessíveis por rede. VMs podem ter unidades de estado sólido (SSDs) anexadas localmente, mas os dados nessas unidades são perdidos quando a VM é terminada. Os dados em discos persistentes continuam a existir após as VMs serem desligadas e terminadas. Discos persistentes existem independentemente das máquinas virtuais; SSDs anexados localmente não.

Características dos Discos Persistentes

Discos persistentes estão disponíveis em configurações SSD e de disco rígido (HDD). SSDs são usados quando alto desempenho é importante. SSDs fornecem desempenho consistente tanto para padrões de acesso aleatório quanto sequencial. HDDs têm latências mais longas, mas custam menos, então HDDs são uma boa opção quando armazenando grandes quantidades de dados e realizando operações em lote que são menos sensíveis à latência do disco do que aplicações interativas. Discos persistentes estão disponíveis nos seguintes tipos:

- Discos persistentes padrão zonais, que fornecem armazenamento de bloco eficiente e confiável dentro de uma zona usando discos rígidos padrão
- Discos persistentes padrão regionais, que são semelhantes aos discos persistentes padrão zonais em desempenho, mas também fornecem replicação síncrona entre duas zonas dentro de uma região
- Discos persistentes balanceados zonais, que são armazenamento custo-efetivo e confiável usando SSDs

- Discos persistentes balanceados regionais, que são semelhantes aos discos persistentes balanceados zonais em desempenho, mas também fornecem replicação síncrona entre duas zonas dentro de uma região
- Discos SSD persistentes zonais, que fornecem armazenamento de bloco rápido e confiável dentro de uma zona
- Discos SSD persistentes regionais, que são semelhantes aos discos SSD persistentes zonais em desempenho, mas também fornecem replicação síncrona entre duas zonas dentro de uma região
- Discos extremos persistentes zonais, que oferecem o armazenamento de bloco de maior desempenho dos discos persistentes e usam SSDs

Além dos discos persistentes, o Google Cloud oferece SSDs Locais, que são armazenamentos locais de bloco de alto desempenho, mas sem redundância. Discos persistentes têm uma capacidade máxima de 64 TB, enquanto os SSDs Locais têm uma capacidade fixa de 375 GB.

Discos persistentes podem ser montados em várias VMs para fornecer armazenamento com múltiplos leitores. Snapshots dos discos podem ser criados em minutos, permitindo que cópias adicionais de dados em um disco sejam distribuídas para uso por outras VMs. Se um disco criado a partir de um snapshot for montado em uma única VM, ele pode suportar operações de leitura e escrita.

O tamanho dos discos persistentes pode ser aumentado enquanto estiverem montados em uma VM. Se você redimensionar um disco, talvez precise executar comandos do sistema operacional para tornar esse espaço adicional acessível ao sistema de arquivos. Tanto discos SSD quanto HDD podem ter até 64 TB.

Discos persistentes criptografam automaticamente os dados no disco.

Ao planejar suas opções de armazenamento, você também deve considerar se deseja que seus discos sejam zonais ou regionais. Discos zonais armazenam dados em vários discos físicos em uma única zona. Se a zona se tornar inacessível, você perderá o acesso aos seus discos. Alternativamente, você pode usar discos persistentes regionais, que replicam blocos de dados em duas zonas dentro de uma região, mas são mais caros do que o armazenamento zonal.

Configurando Discos Persistentes

Você pode criar e configurar discos persistentes a partir do console, navegando até o Compute Engine e selecionando Discos. Na página de Discos, clique em Criar Disco para exibir um formulário como o da Figura 11.2.

Você precisará fornecer um nome para o disco, mas a descrição é opcional. Existem dois tipos de disco: disco persistente padrão e SSD. Para uma maior disponibilidade, você pode ter uma réplica criada dentro da região. Você precisará especificar uma região e zona. Rótulos são opcionais, mas recomendados para ajudar a acompanhar a finalidade de cada disco.

Discos persistentes podem ser criados em branco ou a partir de uma imagem ou snapshot. Use a opção de imagem se quiser criar um disco de boot persistente. Use um snapshot se quiser criar uma réplica de outro disco.

Quando você armazena dados em repouso no Google Cloud, eles são criptografados por padrão. Ao criar um disco, você pode optar por ter o Google gerenciando as chaves de criptografia, caso em que nenhuma configuração adicional é necessária. Você pode usar o Cloud Key Management Service do Google Cloud para gerenciar as chaves por conta própria e armazená-las no repositório de chaves do Google Cloud. Escolha a opção de chave de criptografia gerenciada pelo cliente (CMEK) para isso. Você precisará especificar o nome de uma chave que você criou no Cloud Key Management Service. Se você criar e gerenciar chaves usando outro sistema de gerenciamento de chaves, selecione a chave de criptografia fornecida pelo cliente (CSEK). Você terá que inserir a chave no formulário se escolher a opção de chave fornecida pelo cliente.

Armazenamento de Objetos

Caches são usados para armazenar quantidades relativamente pequenas de dados que devem ser acessíveis com latência submilissegunda. Dispositivos de armazenamento persistente podem armazenar até 64 TB em um único disco e fornecer até centenas de IOPS para operações de leitura e escrita. Quando você precisa armazenar grandes volumes de dados — ou seja, até exabytes — e compartilhá-los amplamente, o armazenamento de objetos é uma boa opção. O armazenamento de objetos do Google

Cloud é o Cloud Storage.

FIGURE 11.2 Form to create a persistent disk

[Create a disk](#)

Name *
disk-1 ?
Name is permanent

Description

Location

Single zone
 Regional
Create a failover replica in the same region for high availability. [Learn more](#)

Region * us-central1 (Iowa) Zone * us-central1-a ?

Source
Create a blank disk, apply a bootable disk image, or restore a snapshot of another disk in this project.

Disk source type * Blank disk

Disk settings

Disk type * Balanced persistent disk ?

[COMPARE DISK TYPES](#)

Size * 100 GB ?
Provision between 10 and 65,536 GB

Snapshot schedule (Recommended)
Use snapshot schedules to automate disk backups. [Learn more](#)

Enable snapshot schedule

Select or create a snapshot schedule * default-schedule-1 ?
Every day, starts between 8:00 AM and 9:00 AM

Encryption
Data is encrypted automatically. Select an encryption key management solution.

Google-managed encryption key
No configuration required

Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

Customer-supplied encryption key (CSEK)
Manage outside of Google Cloud

Actions

[CREATE](#) [CANCEL](#) [EQUIVALENT COMMAND LINE](#) ▾

Características do Cloud Storage

O Cloud Storage é um sistema de armazenamento de objetos, o que significa que os arquivos armazenados no sistema são tratados como unidades atômicas — isto é, você não pode operar em parte do arquivo, como ler apenas uma seção do arquivo. Você pode realizar operações em um objeto, como criar ou deletar, mas o Cloud Storage não oferece funcionalidade para manipular subcomponentes de um arquivo. Por exemplo, não há um comando do Cloud Storage para sobreescriver uma seção do arquivo. Além disso, o Cloud Storage não suporta concorrência e bloqueio. Se múltiplos clientes estiverem escrevendo em um arquivo, então os últimos dados escritos no arquivo são armazenados e persistidos.

O Cloud Storage é bem adequado para armazenar grandes volumes de dados sem requerer qualquer estrutura de dados consistente. Você pode armazenar diferentes tipos de dados em um bucket, que é a unidade lógica de organização no Cloud Storage. Buckets são recursos dentro de um projeto. É importante lembrar que os buckets compartilham um namespace global, então cada nome de bucket deve ser globalmente único. Não deveríamos nos surpreender se não pudermos nomear um bucket como “mytestbucket”, mas não é muito difícil encontrar um nome de arquivo único.

É importante lembrar que o armazenamento de objetos não fornece um sistema de arquivos. Buckets são análogos a diretórios na medida em que ajudam a organizar objetos em grupos, mas buckets não são verdadeiros diretórios que suportam recursos como subdiretórios. O Google suporta um projeto de código aberto chamado Cloud Storage Fuse, que fornece uma maneira de montar um bucket como um sistema de arquivos em sistemas operacionais Linux e Mac. Usando o Cloud Storage Fuse, você pode baixar e subir arquivos para buckets usando comandos de sistema de arquivos, mas ele não fornece funcionalidade completa de sistema de arquivos. O Cloud Storage Fuse tem as mesmas limitações que o Cloud Storage. Seu propósito é tornar mais conveniente mover dados para dentro e fora de buckets quando trabalhando em um sistema de arquivos Linux ou Mac.

O Cloud Storage oferece quatro classes diferentes de armazenamento de objetos: padrão, nearline, coldline e arquivo. Para cada classe de armazenamento, podemos escolher armazenar os dados em uma única região, regiões duplas ou multi-regiões.

Classes de Armazenamento

O armazenamento padrão é a melhor opção para dados frequentemente usados, que às vezes é referido como “dados quentes” ou dados que estão sendo armazenados por curtos períodos de tempo. A replicação em região dupla pode aumentar a disponibilidade em comparação ao armazenamento em região única. O armazenamento em multi-regiões é uma boa opção quando os dados serão lidos de múltiplas regiões e você quer reduzir a latência para acessar dados de várias regiões. O armazenamento padrão em regiões duplas e multi-regiões tem 99,95% de disponibilidade, enquanto a região única tem 99,9% de disponibilidade.

Para dados acessados infrequentemente, as classes de armazenamento nearline e coldline são boas opções. O armazenamento nearline é projetado para casos de uso nos quais você espera acessar arquivos menos de uma vez por mês. O armazenamento coldline é projetado, e especificado, para arquivos esperados para serem acessados uma vez a cada 90 dias ou menos.

O armazenamento nearline tem uma disponibilidade mensal típica de 99,95% em locais multirregionais e uma disponibilidade típica de 99,9% em locais regionais. Os SLAs para nearline são de 99,9% em locais multirregionais e 99,0% em locais regionais. Esses SLAs mais baixos vêm com um custo significativamente menor por gigabyte armazenado, mas antes de começar a mover

antes de mover todos os seus dados regionais e multirregionais para nearline para economizar nos custos, você deve saber que o Google adiciona uma cobrança de recuperação de dados ao armazenamento nearline e coldline. Também há uma duração mínima de armazenamento de 30 dias para o armazenamento nearline.

O armazenamento coldline tem uma disponibilidade mensal típica de 99,95% em locais multirregionais e uma disponibilidade típica de 99,9% em locais regionais. Os SLAs são de 99,9% para locais multirregionais e 99,0% para locais regionais. O coldline também tem um custo por gigabyte menor que o armazenamento nearline. Lembre-se, isso é apenas a cobrança de armazenamento. Como o armazenamento nearline, o armazenamento coldline tem cobranças de acesso. O Google espera que os dados no armazenamento coldline sejam acessados uma vez a cada 90 dias ou menos e tenham um mínimo de armazenamento de 90 dias.

O armazenamento de arquivos é projetado para armazenamento de longo prazo para arquivamento, recuperação de desastres e outros casos de uso onde os dados serão acessados menos de uma vez por ano e serão armazenados por pelo menos 365 dias. O SLA para armazenamento de arquivo é de 99,9% para multi-região e região dupla e 99,0% para tipos de localização de região.

É mais importante entender as relações de custo relativas do que os preços atuais. Os preços podem mudar, mas os custos de cada classe em relação a outras classes de armazenamento são mais propensos a permanecer os mesmos.

Armazenamento Regional, Dual Regional e Multirregional

Quando você cria um bucket, especifica um local para criar o bucket. O bucket e seu conteúdo são armazenados nesta localização. Você pode armazenar seus dados em uma única região, regiões duplas ou múltiplas regiões. Uma região é um local geográfico específico, como Virginia do Norte, Paris e Mumbai. Uma região dupla é um par de regiões. Uma região multirregional é uma grande área geográfica, como os Estados Unidos, União Europeia e Ásia. O SLA de disponibilidade para armazenamento regional é de 99,9%, enquanto a região dupla e multirregional têm uma disponibilidade de 99,95% SLQ. Buckets regionais são redundantes em zonas.

Buckets multirregionais são usados quando o conteúdo precisa ser armazenado em várias regiões para garantir tempos aceitáveis de acesso ao conteúdo. Isso também fornece redundância em caso de falhas em nível de zona. Esses benefícios vêm com um custo mais alto, no entanto. (Não é provável que sejam feitas perguntas sobre preços específicos no exame de Engenheiro de Nuvem Associado, mas você deve conhecer os custos relativos para que possa identificar a solução de menor custo que atenda a um conjunto de requisitos.)

Tanto o armazenamento regional quanto o multirregional são usados para dados geralmente usados. Se você tem uma aplicação onde os usuários baixam e acessam arquivos frequentemente, como mais de uma vez por mês, então é mais custo-efetivo escolher regional ou multirregional. Você escolhe entre regional e multirregional com base na localização dos seus usuários. Se os usuários estão dispersos globalmente e requerem acesso a dados sincronizados, então o multirregional pode fornecer melhor desempenho e disponibilidade.

Uma nota sobre terminologia: O Google às vezes usa o termo georreduntante. Dados georreduntantes são armazenados em pelo menos dois locais que estão a pelo menos 100 milhas de distância. Se seus dados estão em locais multirregionais, então eles são georreduntantes.

Versionamento e Gerenciamento do Ciclo de Vida de Objetos

Buckets no Cloud Storage podem ser configurados para reter versões de objetos quando eles são alterados. Quando o versionamento é ativado em um bucket, uma cópia de um objeto é arquivada cada vez que o objeto é sobreescrito ou quando é excluído. A versão mais recente do objeto é conhecida como a versão ativa. O versionamento é útil quando você precisa manter um histórico de alterações em um objeto ou quer mitigar o risco de excluir acidentalmente um objeto.

O Cloud Storage também oferece políticas de gerenciamento de ciclo de vida para alterar automaticamente a classe de armazenamento de um objeto ou excluir o objeto após um período especificado. Uma política de ciclo de vida, às vezes chamada de configuração, é um conjunto de regras. As regras incluem uma condição e uma ação. Se a condição for verdadeira, a ação é executada. As políticas de gerenciamento de ciclo de vida são aplicadas a buckets e afetam todos os objetos no bucket.

As condições são frequentemente baseadas na idade. Uma vez que um objeto atinge uma certa idade, ele pode ser excluído ou movido para uma classe de armazenamento de menor custo. Além da idade, as condições podem verificar o número de versões, se a versão está ativa, se o objeto foi criado antes de uma data específica e se o objeto está em uma classe de armazenamento particular.

Você pode excluir um objeto ou alterar sua classe de armazenamento. Tanto objetos não versionados quanto versionados podem ser excluídos. Se a versão ativa de um arquivo for excluída, em vez de realmente excluí-la, o objeto é arquivado. Se uma versão arquivada de um objeto for excluída, o objeto é permanentemente excluído.

Você também pode alterar a classe de armazenamento de um objeto usando o gerenciamento de ciclo de vida. Existem restrições sobre quais classes podem ser atribuídas. Objetos de armazenamento padrão podem ser alterados para nearline, coldline ou arquivo. Nearline pode ser alterado apenas para coldline ou arquivo, enquanto coldline pode ser alterado para arquivo.

Configurando o Cloud Storage

Você pode criar buckets no Cloud Storage usando o console. A partir do menu principal, navegue até Armazenamento e selecione Criar Bucket. Isso exibirá um formulário semelhante à Figura 11.3.

Ao criar um bucket, você precisa fornecer algumas informações básicas, incluindo um nome de bucket e classe de armazenamento. Você pode opcionalmente adicionar rótulos e escolher entre chaves gerenciadas pelo Google ou chaves gerenciadas pelo cliente para criptografia. Você também pode definir uma política de retenção para evitar mudanças nos arquivos ou exclusão de arquivos antes do tempo que você especificar.

Uma vez que você criou um bucket, você define uma política de ciclo de vida. Escolha Ciclo de Vida no menu horizontal para exibir o formulário mostrado na Figura 11.4.

Observe que a coluna Ciclo de Vida indica se uma configuração de ciclo de vida está habilitada. Escolha um bucket para criar ou modificar um ciclo de vida e clique em Nenhum ou Habilitado na coluna Ciclo de Vida. Isso exibirá o formulário mostrado na Figura 11.5.

FIGURE 11.3 Form to create a storage bucket from the console. Advanced options are displayed.

The screenshot shows the 'Create a bucket' interface. At the top left is a back arrow and the title 'Create a bucket'. The main area contains several sections:

- Name your bucket**: A text input field with placeholder 'Ex: 'example', 'example_bucket-1', or 'example.com''. Below it is a tip: 'Tip: Don't include any sensitive information'.
- LABELS (OPTIONAL)**: A section with a 'CONTINUE' button.
- Choose where to store your data**: Shows 'Location: us (multiple regions in United States)' and 'Location type: Multi-region'.
- Choose a default storage class for your data**: Shows 'Default storage class: Standard'.
- Choose how to control access to objects**: Shows 'Public access prevention: Off' and 'Access control: Uniform'.
- Choose how to protect object data**: Shows 'Protection tools: None' and 'Data encryption: Google-managed key'.

To the right, under 'Good to know', there's a section titled 'Location pricing' with a note about varying storage rates based on location. It also shows the current configuration as 'Multi-region / Standard' and provides an estimate of monthly costs for the 'us' region at '\$0.026 per GB-month'.

At the bottom are 'CREATE' and 'CANCEL' buttons.

Quando você adiciona uma regra, precisa especificar a condição do objeto e a ação. As opções de condição são Idade, Data de Criação, Classe de Armazenamento, Versões Mais Novas e Estado Ativo. O Estado Ativo aplica-se a objetos versionados, e você pode configurar sua condição para aplicar-se a versões ao vivo ou arquivadas de um objeto. A ação pode definir a classe de armazenamento para nearline, coldline ou arquivo.

Vamos olhar para um exemplo de política. Na seção Navegador do Cloud Storage no console, você pode ver uma lista de buckets, como mostrado na Figura 11.6.

FIGURE 11.4 The list of buckets includes a link to define or modify life cycle policies.

The screenshot shows a table of buckets. One bucket is listed:

Location	Storage class	Public access	Protection
us-west1 (Oregon)	Standard	Not public	None

Below the table, there's a section titled "LIFECYCLE" with the following text:

Lifecycle rules let you apply actions to a bucket's objects when certain conditions are met – for example, switching objects to colder storage classes when they reach or pass a certain age. [Learn more](#)

If an object meets the conditions for multiple rules:

- Deletion takes precedence over a change in storage class.
- Changing objects to colder storage classes takes precedence over changing to warmer ones (ex. objects will switch to the Archive storage class instead of Coldline if there are rules for both).

Below this, there's a "Rules" section with "ADD A RULE" and "DELETE ALL" buttons. A table below shows columns for Action, Object condition, and Works with. The message "You haven't added any lifecycle rules to this bucket." is displayed.

Tipos de Armazenamento ao Planejar uma Solução de Armazenamento

Ao planejar uma solução de armazenamento, um fator que você deve considerar é o tempo necessário para acessar dados. Caches, como Memorystore, oferecem o tempo de acesso mais rápido, mas são limitados à quantidade de memória disponível. Caches são voláteis; quando o servidor é desligado, os conteúdos do cache são perdidos. Você deve salvar os conteúdos do cache em armazenamento persistente em intervalos regulares para permitir a recuperação ao ponto no tempo em que os conteúdos do cache foram salvos pela última vez.

O armazenamento persistente é usado para dispositivos de armazenamento de bloco, como discos anexados a VMs. O Google Cloud oferece drives SSD e HDD. SSDs fornecem desempenho mais rápido, mas custam mais. HDDs são usados quando grandes volumes de dados precisam ser armazenados em um sistema de arquivos, mas os usuários dos dados não precisam do acesso mais rápido possível.

O armazenamento de objetos é usado para armazenar grandes volumes de dados por longos períodos de tempo. O Cloud Storage tem classes de armazenamento regional e multiregional e suporta gerenciamento de ciclo de vida e versionamento.

Além de escolher um sistema de armazenamento subjacente, você terá que considerar como os dados são armazenados e acessados. Para isso, é importante entender os modelos de dados disponíveis e quando usá-los.

FIGURE 11.5 When creating a life cycle policy, click the Add Rule option, which is in the lower horizontal menu, to define a rule.

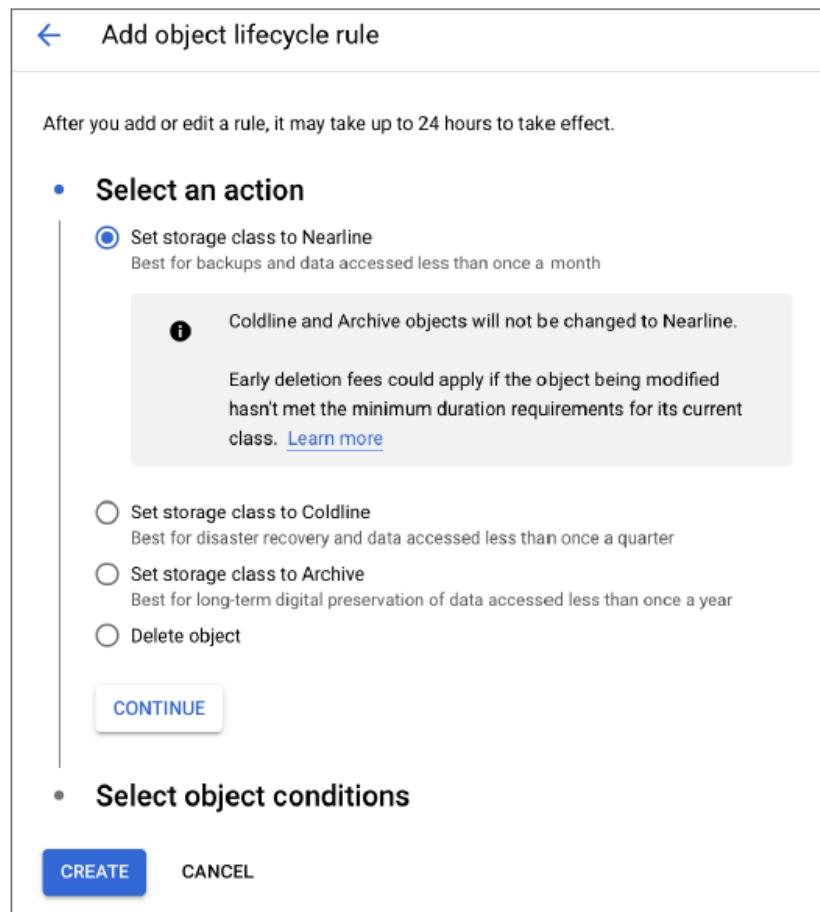


FIGURE 11.6 Listing of buckets in Cloud Storage Browser

Filter <input type="text" value="Filter buckets"/>					
<input type="checkbox"/>	Name	Created	Location type	Location	Default storage class
<input type="checkbox"/>	dataflow-staging-us-west1-38894734...	Oct 23, 2022, 11:06:45 AM	Region	us-west1	Standard
<input type="checkbox"/>	gcf-sources-388947348090-us-central1	Nov 19, 2022, 9:52:45 AM	Region	us-central1	Standard
<input type="checkbox"/>	slg-cloud-storage-2	Oct 26, 2022, 6:55:12 AM	Region	us-west1	Standard

Modelos de Dados de Armazenamento

Existem quatro categorias amplas de modelos de dados disponíveis no Google Cloud: objeto, relacional, analítico e NoSQL.

Objeto: Cloud Storage

O modelo de dados de armazenamento de objetos trata arquivos como objetos atômicos. Você não pode usar comandos de armazenamento de objetos para ler blocos de dados ou sobreescriver partes do objeto. Se você precisar atualizar um objeto, deve copiá-lo para um servidor, fazer a alteração e, em seguida, copiar a versão atualizada de volta para o sistema de armazenamento de objetos.

O armazenamento de objetos é usado quando você precisa armazenar grandes volumes de dados e não precisa de acesso granular aos dados dentro de um objeto enquanto ele está no armazenamento de objetos. Este modelo de dados é bem adequado para dados arquivados, dados de treinamento de aprendizado de máquina e dados antigos da Internet das Coisas (IoT) que precisam ser salvos, mas não são mais ativamente analisados.

Relacional: Cloud SQL e Cloud Spanner

Bancos de dados relacionais têm sido o principal armazenamento de dados para empresas por décadas. Bancos de dados relacionais suportam consultas e atualizações frequentes de dados. Eles são usados quando é importante que os usuários tenham uma visão consistente dos dados. Por exemplo, se dois usuários estão lendo dados de uma tabela relacional ao mesmo tempo, eles verão os mesmos dados. Isso nem sempre é o caso com bancos de dados que podem ter inconsistências entre réplicas de dados, como alguns bancos de dados NoSQL.

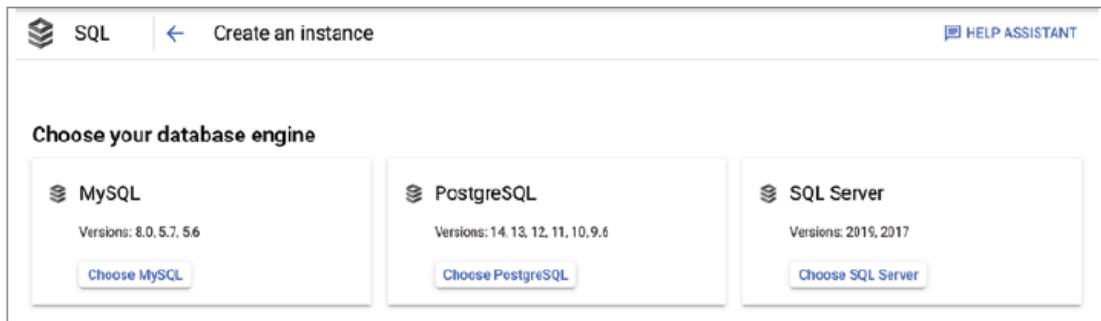
Bancos de dados relacionais, como Cloud SQL e Cloud Spanner, suportam transações de banco de dados. Uma transação é um conjunto de operações garantidas para ter sucesso ou falhar em sua totalidade — não há chance de que algumas operações sejam executadas e outras não. Por exemplo, quando um cliente compra um produto, a contagem do número de produtos disponíveis é decrementada na tabela de inventário e um registro é adicionado a uma tabela de produtos comprados pelo cliente. Com transações, se o banco de dados falhar após atualizar o inventário, mas antes de atualizar a tabela de produtos comprados pelo cliente, o banco de dados reverterá a transação parcialmente executada quando o banco de dados reiniciar.

Cloud SQL e Cloud Spanner são usados quando os dados são estruturados e modelados para bancos de dados relacionais. Cloud SQL é um serviço de banco de dados gerenciado que fornece bancos de dados MySQL, SQL Server e PostgreSQL. Cloud SQL é usado para bancos de dados que não precisam escalar horizontalmente — ou seja, adicionando servidores adicionais a um cluster. Bancos de dados Cloud SQL escalam verticalmente — ou seja, rodando em servidores com mais memória e mais CPU. Cloud Spanner é usado quando você tem volumes extremamente grandes de dados relacionais ou dados que precisam ser distribuídos globalmente, garantindo consistência e integridade de transação em todos os servidores. Grandes empresas costumam usar Cloud Spanner para aplicações como cadeias de suprimentos globais e aplicações de serviços financeiros, enquanto Cloud SQL é frequentemente usado para aplicações web e aplicações de e-commerce.

Configurando Cloud SQL

Você pode criar uma instância Cloud SQL navegando até Cloud SQL no menu principal do console e selecionando Criar Instância. Você será solicitado a escolher uma instância MySQL, PostgreSQL ou SQL Server, como mostrado na Figura 11.7.

FIGURE 11.7 Cloud SQL provides MySQL, PostgreSQL, and SQL Server instances.



Para configurar uma instância MySQL, você precisará especificar um nome, senha root, região e zona. As opções de configuração incluem o seguinte:

- Versão do MySQL.
- Conectividade, onde você pode especificar se deseja usar um endereço IP público ou privado.
- Tipo de máquina. O padrão é um db-n1-standard-1 com 1 vCPU e 3.75 GB de memória.
- Backups automáticos.
- Réplicas de failover.
- Flags de banco de dados. Estes são específicos para o MySQL e incluem a capacidade de definir uma flag de somente leitura para o banco de dados e definir o tamanho do cache de consulta.
- Uma janela de tempo de manutenção.
- Rótulos.

Figura 11.8 mostra o formulário de configuração para instâncias MySQL, Figura 11.9 mostra a configuração para instâncias SQL Server, e Figura 11.10 mostra a configuração para instâncias PostgreSQL.

Configurando Cloud Spanner

Se você precisar criar um banco de dados global e consistente com suporte para transações, você deve considerar Cloud Spanner. Dada a natureza avançada do Spanner, sua configuração é surpreendentemente simples. No console, navegue até Cloud Spanner e selecione Criar Instância para exibir o formulário na Figura 11.11.

FIGURE 11.8 Configuration form for a MySQL instance

The screenshot shows the 'Create a MySQL Instance' configuration interface. It includes the following sections:

- Instance info:** Fields for Instance ID (placeholder: "us-central1"), Password (with GENERATE button), and Database version (MySQL 8.0).
- Choose region and zonal availability:** A note about performance and region permanence. It includes a Region dropdown set to "us-central1 (Iowa)" and a Zonal availability section with two options: "Single zone" (not recommended) and "Multiple zones (Highly available)" (selected). A "SPECIFY ZONES" link is also present.
- Customize your instance:** A note about customizable configurations later, followed by a "SHOW CONFIGURATION OPTIONS" link.
- Summary:** A table showing instance details:

Region	us-central1 (Iowa)
DB Version	MySQL 8.0
vCPUs	4 vCPU
Memory	26 GB
Storage	100 GB
Network throughput (MB/s)	1,000 of 2,000
Disk throughput (MB/s)	Read: 48.0 of 240.0 Write: 48.0 of 240.0
IOPS	Read: 3,000 of 15,000 Write: 3,000 of 15,000
Connections	Public IP
Backup	Automated
Availability	Multiple zones (Highly available)
Point-in-time recovery	Enabled

At the bottom are "CREATE INSTANCE" and "CANCEL" buttons.

Você precisa fornecer um nome de instância, ID de instância e número de nós. Você também terá que escolher entre uma configuração regional ou multirregional para determinar onde os nós e dados estão localizados. Isso determinará o custo e a localização do armazenamento de replicação. Se você selecionar Regional, escolherá da lista de regiões disponíveis, como us-west1, asia-east1 e europe-north1.

Analítico: BigQuery

O BigQuery é um serviço projetado para um data warehouse e aplicações analíticas. O BigQuery é projetado para armazenar petabytes de dados. O BigQuery trabalha com grandes números de linhas e colunas de dados e não é adequado para aplicações orientadas a transações, como e-commerce ou suporte para aplicações web interativas.

FIGURE 11.9 Configuration form for a SQL Server instance

The screenshot shows the 'Create a SQL Server instance' configuration form. It consists of several sections:

- Instance info:** Fields for 'Instance ID' (with placeholder 'Use lowercase letters, numbers, and hyphens. Start with a letter.'), 'Password' (with a 'GENERATE' button and note 'Your default service admin username is "sqlserver"'), and 'Database version' (set to 'SQL Server 2019 Standard').
- Choose region and zonal availability:** A note 'For better performance, keep your data close to the services that need it. Region is permanent, while zone can be changed any time.' Below this, the 'Region' is set to 'us-central1 (Iowa)'. Under 'Zonal availability', the 'Multiple zones (Highly available)' option is selected (indicated by a blue circle). A note states: 'Automatic failover to another zone within your selected region. Recommended for production instances. Increases cost.' There is also a link to 'SPECIFY ZONES'.
- Customize your instance:** A note 'You can also customize instance configurations later' and a link to 'SHOW CONFIGURATION OPTIONS'.
- Summary:** A table showing instance specifications:

Region	us-central1 (Iowa)
DB Version	SQL Server 2019 Standard
vCPUs	4 vCPU
Memory	26 GB
Storage	100 GB
Network throughput (MB/s)	1,000 of 2,000
Disk throughput (MB/s)	Read: 48.0 of 240.0 Write: 48.0 of 240.0
IOPS	Read: 3,000 of 15,000 Write: 3,000 of 15,000
Connections	Public IP
Backup	Automated
Availability	Multiple zones (Highly available)

At the bottom are 'CREATE INSTANCE' and 'CANCEL' buttons.

Configurando BigQuery

O BigQuery é um serviço de análise sem servidor, que fornece armazenamento além de ferramentas de consulta, análise estatística e aprendizado de máquina. O BigQuery não requer que você configure instâncias. Em vez disso, quando você navega pela primeira vez até o BigQuery a partir do menu do console, você verá o formulário mostrado na Figura 11.12.

A primeira tarefa para usar o BigQuery é criar um conjunto de dados para armazenar dados. Você faz isso clicando em Criar Conjunto de Dados para exibir o formulário mostrado na Figura 11.13.

Ao criar um conjunto de dados, você terá que especificar um nome e selecionar uma região onde armazená-lo. Nem todas as regiões suportam o BigQuery. Atualmente, você tem a escolha de muitos locais nos Estados Unidos, Europa e Ásia.

FIGURE 11.10 Configuration form for a PostgreSQL instance

The screenshot shows the 'Create a PostgreSQL instance' configuration page. It includes sections for 'Instance info', 'Choose region and zonal availability', 'Customize your instance', and a 'Summary' sidebar.

Instance info:

- Instance ID:
- >Password: [GENERATE](#)
- Database version: PostgreSQL 14

Choose region and zonal availability:

For better performance, keep your data close to the services that need it. Region is permanent, while zone can be changed any time.

Region: us-central1 (Iowa)

Zonal availability:

- Single zone: In case of outage, no failover. Not recommended for production.
- Multiple zones (Highly available): Automatic failover to another zone within your selected region. Recommended for production instances. Increases cost.

Customize your instance:

You can also customize instance configurations later

Summary:

Region	us-central1 (Iowa)
DB Version	PostgreSQL 14
vCPUs	4 vCPU
Memory	26 GB
Storage	100 GB
Network throughput (MB/s)	1,000 of 2,000
Disk throughput (MB/s)	Read: 48.0 of 240.0 Write: 48.0 of 240.0
IOPS	Read: 3,000 of 15,000 Write: 3,000 of 15,000
Connections	Public IP
Backup	Automated
Availability	Multiple zones (Highly available)
Point-in-time recovery	Enabled

Buttons:

- [CREATE INSTANCE](#)
- [CANCEL](#)

No Capítulo 12, discutiremos como carregar e consultar dados no BigQuery e em outros bancos de dados do Google Cloud.

NoSQL: Cloud Firestore e Bigtable

Bancos de dados NoSQL não usam o modelo relacional e não requerem uma estrutura ou esquema fixo. Esquemas de banco de dados definem quais tipos de atributos podem ser armazenados. Quando nenhum esquema fixo é necessário, os desenvolvedores têm a opção de armazenar diferentes atributos em registros diferentes. O Google Cloud possui um banco de dados de documentos chamado Cloud Firestore e um banco de dados de coluna ampla chamado Bigtable.

FIGURE 11.11 The Cloud Spanner configuration form in Cloud Console

The screenshot shows the 'Create an instance' configuration form in the Google Cloud Console. It includes sections for 'Name your instance', 'Choose a configuration', and 'Allocate compute capacity'. The 'Summary' section on the right provides an overview of storage and compute costs.

Section	Setting	Description
Name your instance	Instance name *	An instance has both a name and an ID. The name is for display purposes only. The ID is a permanent and unique identifier.
	Instance ID *	Name must be 4-30 characters long. Lowercase letters, numbers, hyphens allowed.
Choose a configuration	Regional (selected)	Determines where your nodes and data are located. Affects cost, performance, and replication. A multi-region configuration will select the default leader region for your leader replicas. You can change your leader region at any time with a DDL statement. Learn more
Allocate compute capacity	Unit * Processing units	Your compute capacity determines the amount of data throughput, queries per second (QPS), and storage limits in your instance. One node equals 1,000 processing units. Affects billing.
	Quantity * 1000	Integers only. Enter in increments of 100 up to 1,000, followed by increments of 1,000.
CREATE		CANCEL

FIGURE 11.12 BigQuery user interface for creating and querying data

The screenshot shows the BigQuery user interface. The left pane is the 'Explorer' showing a search bar and a list of resources under 'scenic-energy-335022'. The right pane is the 'Editor' with a query editor area containing the number '1' and a status bar with 'Type a query to get started'.

FIGURE 11.13 Form to create a data set in BigQuery

The screenshot shows the 'Create dataset' interface. At the top, it says 'Create dataset'. Below that, 'Project ID' is set to 'scenic-energy-335022' with a 'CHANGE' button. A 'Dataset ID' field is marked with a red asterisk, indicating it is required. Below it, a note says 'Letters, numbers, and underscores allowed'. A 'Data location' dropdown is shown. Under 'Default table expiration', there is a checkbox for 'Enable table expiration' with a help icon. A 'Default maximum table age' field with a 'Days' unit is also present. An 'Advanced options' section is collapsed. At the bottom are 'CREATE DATASET' and 'CANCEL' buttons.

Recursos do Firestore

Firestore é um banco de dados de documentos. Isso não significa que ele é usado para armazenar documentos como planilhas ou arquivos de texto, mas que os dados no banco de dados são organizados em uma estrutura chamada documento. Documentos são compostos por conjuntos de pares chave-valor. Um exemplo simples é o seguinte:

```
{  
    livro: "Guia de Estudo ACE",  
    capítulo: 11,  
    comprimento: 20,  
    tópico: "armazenamento"  
}
```

Este exemplo descreve as características de um capítulo em um livro. Há quatro chaves ou propriedades neste exemplo: livro, capítulo, comprimento e tópico. Este conjunto de pares chave-valor é chamado de entidade na terminologia do Firestore. Entidades frequentemente têm propriedades em comum, mas, como o Firestore é um banco de dados sem esquema, não há requisito de que todas as entidades tenham o mesmo conjunto de propriedades. Aqui está um exemplo:

```
{
```

```
livro: "Guia de Estudo ACE",
capítulo: 11,
tópico: "computação",
número_de_figuras: 8
}
```

O Firestore é um banco de dados gerenciado, então os usuários do serviço não precisam gerenciar servidores ou instalar software de banco de dados. O Firestore partitiona dados automaticamente e escala para cima ou para baixo conforme a demanda justificar.

O Firestore é usado para necessidades de armazenamento não analíticas e não relacionais. É uma boa escolha para catálogos de produtos, que têm muitos tipos de produtos com características ou propriedades variáveis. Também é uma boa escolha para armazenar perfis de usuários associados a uma aplicação.

O Firestore tem alguns recursos em comum com bancos de dados relacionais, como suporte para transações e índices para melhorar o desempenho da consulta. A principal diferença é que o Firestore não requer um esquema ou estrutura fixa e não suporta operações relacionais, como a junção de tabelas, ou cálculo de agregações, como somas e contagens.

O Cloud Firestore é a última geração de bancos de dados de documentos no Google Cloud. O Cloud Datastore precedeu o Cloud Firestore como um banco de dados de documentos.

Configurando Firestore

Firestore, assim como o BigQuery, é um serviço de banco de dados sem servidor que não requer que você especifique configurações de nó. Em vez disso, você pode trabalhar a partir do console para adicionar entidades ao banco de dados. A Figura 11.14 mostra o formulário inicial que aparece quando você navega pela primeira vez até o Firestore no Cloud Console. A primeira coisa que você deve fazer ao usar o Firestore é escolher entre o modo Nativo, que escala automaticamente para milhões de clientes, ou o modo Datastore, que escala automaticamente para milhões de escritas por segundo.

Depois de escolher um modo, você escolhe onde armazenar seus dados (veja a Figura 11.15). Você tem a opção de usar armazenamento multirregional ou armazenamento regional.

Uma vez que você configurou o Cloud Firestore para o seu projeto no modo Datastore, você pode criar entidades. Ao criar uma entidade, você especifica um namespace, que é uma maneira de agrupar entidades assim como esquemas agrupam tabelas em um banco de dados relacional. Você precisará especificar um tipo, que é análogo a uma tabela em um banco de dados relacional. Cada entidade requer uma chave, que pode ser uma chave numérica gerada automaticamente ou uma chave definida personalizadamente.

Em seguida, você adicionará uma ou mais propriedades que têm nomes, tipos e valores. Os tipos incluem string, data e hora, Booleano e outros tipos estruturados como arrays.

O modo Nativo do Firestore oferece um modelo de dados diferente baseado em documentos e coleções. Documentos são coleções de pares chave-valor e coleções são conjuntos de documentos. Detalhes adicionais sobre carregar e consultar dados no Firestore estão no Capítulo 12.

Recursos do Bigtable

Bigtable é outro banco de dados NoSQL, mas, ao contrário do Firestore, é um banco de dados de coluna ampla, não um banco de dados de documentos. Bancos de dados de coluna ampla, como o nome indica, armazenam tabelas que podem ter um grande número de colunas. Nem todas as linhas precisam usar todas as colunas, então dessa forma é como o Firestore — nenhum dos dois requer um esquema fixo para estruturar os dados.

FIGURE 11.14 The Firestore user interface allows you to choose between Native and Datastore modes.

1 Select a Cloud Firestore mode — 2 Choose where to store your data

Cloud Firestore is the next generation of Cloud Datastore. You can use Cloud Firestore in either Native mode or Datastore mode, each with distinct system behavior optimized for different types of projects. [Pricing](#) for both modes is based on location, stored data, operations, and network egress with a daily free quota for each. [Learn more about choosing a mode](#)

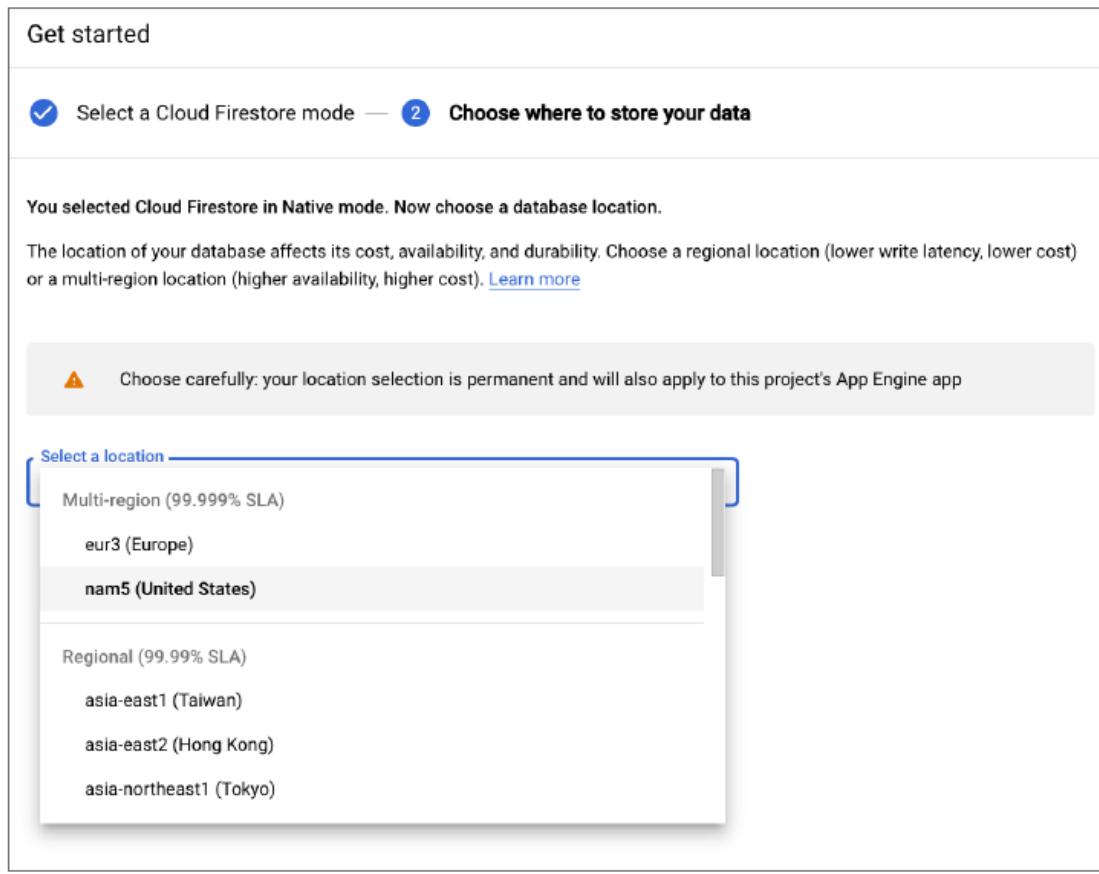
⚠️ The mode you select here will be permanent for this project

	Native mode	Datastore mode
API	Firestore	Datastore
Scalability	Automatically scales to millions of concurrent clients	Automatically scales to millions of writes per second
App engine support	Not supported in the App Engine standard Python 2.7 and PHP 5.5 runtimes	All runtimes
Max writes per second	10,000	No limit
Real-time updates	✓	✗
Mobile/web client libraries with offline data persistence	✓	✗

Bigtable é projetado para bancos de dados de escala de petabyte. Tanto bancos de dados operacionais, como armazenar dados IoT, quanto processamento analítico, como aplicações de ciência de dados, podem usar efetivamente o Bigtable. Este banco de dados é projetado para fornecer latência consistente e baixa em milissegundos. O Bigtable é executado em clusters e escala horizontalmente.

O Bigtable é projetado para aplicações com altos volumes de dados e ingestão de alta velocidade de dados. Séries temporais, IoT e aplicações financeiras se enquadram nesta categoria.

FIGURE 11.15 Choosing a storage location



Configurando Bigtable

A partir do Cloud Console, navegue até o Bigtable e clique em Criar Instância para abrir o formulário mostrado na Figura 11.16.

Neste formulário, você precisará fornecer um nome de instância e um ID de instância. Em seguida, escolha entre o modo Produção ou Desenvolvimento. Clusters de produção têm um mínimo de três nós e fornecem alta disponibilidade. O modo Desenvolvimento usa instâncias de baixo custo sem replicação ou alta disponibilidade. Você também precisará escolher entre SSD ou HDD para discos persistentes usados pelo banco de dados.

O Bigtable pode suportar vários clusters. Para cada cluster, você precisará especificar um ID de cluster, uma localização de região e zona, e o número de nós no cluster. O cluster pode ser replicado para melhorar a disponibilidade.

No Capítulo 12, descreveremos como carregar e consultar dados no Bigtable.

FIGURE 11.16 Configuration form for Bigtable

The screenshot shows the 'Create an instance' configuration form for Google Bigtable. At the top, there's a back arrow labeled 'Create an instance'. Below it, a note says 'A Bigtable instance is a container for your clusters. [Learn more](#)'. To the right, the estimated monthly cost is listed as '\$468 per month (estimated)' with a note that 'That's about \$0.65 an hour with 0 GB stored.' A 'SHOW DETAILS' button is also present.

The form is divided into three steps:

- 1 Name your instance**: Fields include 'Instance name*' (with placeholder 'For display purposes only') and 'Instance ID*' (with placeholder 'ID is permanent'). A 'CONTINUE' button is below these fields.
- 2 Select your storage type**
- 3 Configure your first cluster**

At the bottom, there are 'SHOW ADVANCED OPTIONS' and 'CREATE' (highlighted in blue) and 'CANCEL' buttons.

Mundo Real

A Necessidade de Múltiplos Bancos de Dados

Organizações de saúde e instalações médicas armazenam e gerenciam uma ampla gama de dados sobre pacientes, seus tratamentos e os resultados. Os registros médicos de um paciente incluem informações demográficas, como nome, endereço, idade e assim por diante. Os registros médicos também armazenam informações detalhadas sobre condições médicas e diagnósticos, bem como tratamentos, como medicamentos prescritos e procedimentos realizados. Esse tipo de dado é altamente estruturado. Suporte a transações e forte consistência são necessários. Bancos de dados relacionais, como o Cloud SQL, são uma boa solução para esse tipo de aplicação.

Os dados médicos armazenados em bancos de dados relacionais transacionais são valiosos para analisar padrões em tratamentos e recuperação. Por exemplo, cientistas de dados podem usar registros médicos para identificar padrões associados à readmissão hospitalar. No entanto, bancos de dados relacionais transacionais não são adequados para análises. Uma opção melhor é usar o BigQuery e construir um data warehouse com dados estruturados de maneiras que tornem mais fácil analisar os dados. Dados do sistema transacional são extraídos, transformados e carregados em um conjunto de dados do BigQuery.

Escolhendo uma Solução de Armazenamento: Diretrizes a Considerar

O Google Cloud oferece várias soluções de armazenamento. Como um engenheiro de nuvem, você pode ter que ajudar a planejar e implementar soluções de armazenamento para uma ampla gama de aplicações. As diferentes soluções de armazenamento se

adequam a diferentes casos de uso e, em muitas aplicações empresariais, você descobrirá que precisa de dois ou mais produtos de armazenamento para suportar toda a gama de requisitos da aplicação. Aqui estão vários fatores a ter em mente ao escolher soluções de armazenamento:

Padrões de Leitura e Escrita: Algumas aplicações, como contabilidade e vendas no varejo, leem e escrevem dados frequentemente. Também há atualizações frequentes nessas aplicações. Elas são melhor atendidas por uma solução de armazenamento como o Cloud SQL se os dados forem estruturados; no entanto, se você precisar de um banco de dados global que suporte operações de leitura/escrita relacionais, então o Cloud Spanner é uma escolha melhor. Se você está escrevendo dados em taxas consistentemente altas e em grandes volumes, considere o Bigtable. Se você está escrevendo arquivos e, em seguida, baixando-os inteiramente, o Cloud Storage é uma boa opção.

Consistência: A consistência garante que um usuário lendo dados do banco de dados obterá os mesmos dados, não importa qual servidor em um cluster responda à solicitação. Se você precisa de forte consistência, que está sempre lendo os dados mais recentes, então o Cloud SQL e o Cloud Spanner são boas opções. O Firestore pode ser configurado para forte consistência, mas as operações de I/O levarão mais tempo do que se uma configuração de consistência menos estrita for usada. O Firestore é uma boa opção se seus dados forem não estruturados; caso contrário, considere um dos bancos de dados relacionais. Bancos de dados NoSQL oferecem pelo menos consistência eventual, o que significa que algumas réplicas podem não estar sincronizadas por um curto período de tempo. Durante esses períodos é possível ler dados desatualizados. Se sua aplicação pode tolerar isso, então você pode descobrir que requisitos de consistência menos estritos podem levar a operações de leitura e escrita mais rápidas.

Suporte a Transações: Se você precisa realizar transações atômicas em sua aplicação, use um banco de dados que as suporte. Você pode ser capaz de implementar suporte a transações em sua aplicação, mas esse código pode ser difícil de desenvolver e manter. Os bancos de dados relacionais, Cloud SQL e Spanner, e Firestore fornecem suporte a transações.

Custo: O custo de usar um sistema de armazenamento específico dependerá da quantidade de dados armazenados, da quantidade de dados recuperados ou escaneados, e das cobranças por unidade do sistema de armazenamento. Se você estiver usando um serviço de armazenamento no qual provisiona VMs, você também terá que levar em conta esse custo.

Latência: Latência é o tempo entre o início de uma operação, como uma solicitação para ler uma linha de dados de um banco de dados, até o momento em que ela é concluída. O Bigtable fornece operações consistentemente de baixa latência em milissegundos. O Spanner pode ter latências mais longas, mas com essas latências mais longas você obtém um banco de dados globalmente consistente e escalável.

Em geral, escolher um armazenamento de dados é sobre fazer trocas. Em um mundo ideal, poderíamos ter um banco de dados globalmente escalável, de baixo custo, de baixa latência e fortemente consistente. Não vivemos em um mundo ideal. Temos que abrir mão de uma ou mais dessas características.

No próximo capítulo, você aprenderá como usar cada uma das soluções de armazenamento descritas aqui, com ênfase em carregar e consultar dados.

Resumo

Ao planejar o armazenamento na nuvem, considere os tipos de sistemas de armazenamento e os tipos de modelos de dados. Os sistemas de armazenamento fornecem o hardware e a estrutura organizacional básica usada para armazenar dados. Os modelos de dados organizam os dados em estruturas lógicas que determinam como os dados são armazenados e consultados dentro de um banco de dados.

Os principais sistemas de armazenamento disponíveis no Google Cloud são Memorystore, um serviço gerenciado de cache, e discos persistentes, que são discos acessíveis por rede para VMs no Compute Engine e Kubernetes Engine. O Cloud Storage é o sistema de armazenamento de objetos do Google Cloud.

Os principais modelos de dados são objeto, relacional e NoSQL. Bancos de dados NoSQL no Google Cloud são subdivididos ainda mais em bancos de dados de documentos e de colunas amplas. O Cloud Storage utiliza um modelo de dados de objeto. O Cloud SQL e o Cloud Spanner usam bancos de dados relacionais para aplicações de processamento de transações. O BigQuery usa um modelo relacional para data warehouses e aplicações analíticas. O Firestore é um banco de dados de documentos. O Bigtable é uma tabela de colunas amplas.

Ao escolher sistemas de armazenamento de dados, considere padrões de leitura e escrita, requisitos de consistência, suporte a transações, custo e latência.

Essenciais para o Exame

Conheça os principais tipos de sistema de armazenamento, incluindo caches, discos persistentes e armazenamento de objetos. Caches são usados para melhorar o desempenho da aplicação reduzindo a necessidade de ler de bancos de dados em disco. Caches são limitados pela quantidade de memória disponível. Discos persistentes são dispositivos de rede que são anexados a VMs. Discos persistentes podem ser anexados a múltiplas VMs em modo somente leitura. O armazenamento de objetos é usado para armazenar arquivos para acesso compartilhado e armazenamento de longo prazo.

Conheça os principais tipos de modelos de dados. Bancos de dados relacionais são usados para sistemas de processamento de transações que requerem suporte a transações e consistência forte. O Cloud SQL e o Cloud Spanner são bancos de dados relacionais usados para aplicações de processamento de transações. O BigQuery usa um modelo analítico, mas é projetado para data warehouses e análises. O modelo de objeto é uma alternativa ao modelo de sistema de arquivos. Objetos, armazenados como arquivos, são tratados como unidades atômicas. Modelos de dados NoSQL incluem modelos de dados de documentos e modelos de colunas amplas. O Firestore é um banco de dados de documentos. O Bigtable é um banco de dados de colunas amplas.

Conheça as várias classes no Cloud Storage. Standard, nearline, coldline e arquivo são as quatro classes de armazenamento. Standard é projetado para dados que são acessados frequentemente (mais de uma vez por mês) ou apenas armazenados no Cloud

Storage por um curto período de tempo. Nearline é projetado para acesso infrequente, menos de uma vez por mês. O armazenamento Coldline é projetado para armazenamento de longo prazo, com arquivos sendo acessados menos de uma vez a cada 90 dias. O armazenamento de arquivo é projetado para dados que não são acessados mais frequentemente do que uma vez por ano. Nearline, Coldline e armazenamento de arquivo incorrem em cobranças de recuperação além das cobranças baseadas no tamanho dos dados.

Saiba que aplicações na nuvem podem requerer mais de um tipo de armazenamento de dados. Por exemplo, uma aplicação pode precisar de um cache para reduzir a latência ao consultar dados no Cloud SQL, armazenamento de objetos para o armazenamento de longo prazo de arquivos de dados, e o BigQuery para relatórios e análises de data warehousing.

Saiba que você pode aplicar configurações de ciclo de vida em buckets do Cloud Storage. Ciclos de vida são usados para excluir arquivos e alterar a classe de armazenamento. Objetos da classe Standard podem ser alterados para Nearline, Coldline ou Arquivo. O armazenamento Nearline pode mudar para Coldline e Arquivo. Coldline pode ser alterado para Arquivo.

Conheça as características de diferentes armazenamentos de dados que ajudam a determinar qual é a melhor opção para seus requisitos. Padrões de leitura e escrita, requisitos de consistência, suporte a transações, custo e latência são frequentemente fatores.

Questões

1. Você foi encarregado de definir configurações de ciclo de vida em buckets no Cloud Storage. Você precisa considerar todas as opções possíveis para transição de uma classe de armazenamento para outra. Todas as transições a seguir são permitidas, exceto qual?
 - A. Nearline para Coldline
 - B. Coldline para Arquivo
 - C. Padrão para Nearline
 - D. Arquivo para Padrão
2. Seu gerente pediu sua ajuda para reduzir as cobranças do Cloud Storage. Você sabe que alguns dos arquivos armazenados no Cloud Storage são raramente acessados mais do que uma vez a cada 90 dias. Que tipo de armazenamento você recomendaria para esses arquivos?
 - A. Nearline
 - B. Padrão
 - C. Coldline
 - D. Arquivo
3. Você está trabalhando com uma startup desenvolvendo software de análise para dados de IoT. Você precisa ingerir grandes volumes de dados consistentemente e armazená-los por vários meses. A startup tem várias aplicações que precisarão consultar esses dados. Espera-se que os volumes cresçam para volumes de petabytes. Qual banco de dados você deveria usar?
 - A. Cloud Spanner
 - B. Bigtable
 - C. BigQuery
 - D. Firestore
4. Um desenvolvedor de software em sua equipe está pedindo sua ajuda para melhorar o desempenho de consulta de uma aplicação de banco de dados. O desenvolvedor está usando um banco de dados Cloud SQL MySQL e está disposto a modificar algumas partes da aplicação, mas quer continuar a usar um banco de dados relacional. Quais opções você recomendaria?
 - A. Memorystore e discos persistentes SSD
 - B. Memorystore e discos persistentes HDD
 - C. Firestore e discos persistentes SSD

- D. Firestore e discos persistentes HDD
5. Você está criando um conjunto de discos persistentes para armazenar dados para análise exploratória de dados. Os discos serão montados em uma máquina virtual na zona us-west2-a. Os dados são dados históricos recuperados do Cloud Storage. Os analistas de dados não precisam de desempenho máximo e estão mais preocupados com custo do que desempenho. Os dados serão armazenados em um banco de dados relacional local. Que tipo de armazenamento você recomendaria?
- A. SSDs
 - B. HDDs
 - C. Firestore
 - D. Bigtable
6. Qual das seguintes afirmações sobre o Cloud Storage não é verdadeira?
- A. Buckets do Cloud Storage podem ter períodos de retenção.
 - B. Configurações de ciclo de vida podem ser usadas para mudar a classe de armazenamento de Arquivo para Padrão.
 - C. O Cloud Storage não fornece acesso em nível de bloco a dados dentro de arquivos armazenados em buckets.
 - D. O Cloud Storage é projetado para alta durabilidade.
7. Ao usar versionamento em um bucket, como é chamada a versão mais recente do objeto?
- A. Versão Ativa
 - B. Versão Principal
 - C. Versão Atual
 - D. Versão Segura
8. Um gerente de produto pediu seu conselho sobre quais serviços de banco de dados podem ser opções para uma nova aplicação. Transações e suporte para dados tabulares são importantes. Idealmente, o banco de dados suportaria ferramentas comuns de consulta. Que bancos de dados você recomendaria que o gerente de produto considerasse?
- A. BigQuery e Spanner
 - B. Cloud SQL e Spanner
 - C. Cloud SQL e Bigtable
 - D. Bigtable e Spanner
9. O serviço Cloud SQL fornece bancos de dados relacionais totalmente gerenciados. Quais dois tipos de bancos de dados estão disponíveis no Cloud SQL?

- A. Oracle e MySQL
 - B. Oracle e PostgreSQL
 - C. PostgreSQL e MySQL
 - D. MySQL e DB2
10. Qual das seguintes configurações do Cloud Spanner teria o custo horário mais alto?
- A. Localizado em us-central1
 - B. Localizado em nam3
 - C. Localizado em us-west1-a
 - D. Localizado em nam-eur-asia1
11. Quais dos seguintes são serviços de banco de dados que não exigem que você especifique informações de configuração para VMs?
- A. Apenas BigQuery
 - B. Apenas Firestore
 - C. Apenas Bigtable
 - D. BigQuery e Firestore
12. Que tipo de modelo de dados é usado pelo Firestore?
- A. Relacional
 - B. Documento
 - C. Coluna ampla
 - D. Grafo
13. Você foi encarregado de criar um data warehouse para sua empresa. Ele deve suportar dezenas de petabytes de dados e usar SQL como linguagem de consulta. Qual serviço de banco de dados gerenciado você escolheria?
- A. BigQuery
 - B. Bigtable
 - C. Cloud SQL
 - D. IBM DB2
14. Uma equipe de desenvolvedores de aplicativos móveis está desenvolvendo um novo aplicativo. Será necessário sincronizar dados entre dispositivos móveis e um banco de dados de back-end. Qual serviço de banco de dados você recomendaria?
- A. BigQuery
 - B. Firestore

- C. Spanner
 - D. Bigtable
15. Um gerente de produto está considerando um novo conjunto de recursos para um aplicativo que exigirá armazenamento adicional. Quais características de armazenamento você sugeriria que o gerente de produto considere?
- A. Apenas padrões de leitura e escrita.
 - B. Apenas custo.
 - C. Apenas consistência e custo.
 - D. Todas são considerações relevantes.
16. Qual é o tamanho máximo de um cache do Memorystore quando usando Redis?
- A. 100 GB
 - B. 300 GB
 - C. 400 GB
 - D. 50 GB
17. Uma vez que um bucket tem sua classe de armazenamento definida para Arquivo, para quais outras classes de armazenamento ele pode fazer transição?
- A. Padrão
 - B. Nearline
 - C. Coldline
 - D. Nenhuma das opções acima
18. Antes de poder começar a armazenar dados no BigQuery, o que você deve criar?
- A. Um conjunto de dados
 - B. Um bucket
 - C. Um disco persistente
 - D. Uma entidade
19. Quais recursos você pode configurar ao executar um banco de dados MySQL no Cloud SQL?
- A. Tipo de máquina
 - B. Janelas de manutenção
 - C. Réplicas de failover
 - D. Todos os itens acima

20. Um colega está se perguntando por que algumas cobranças de armazenamento são tão altas. Eles explicam que moveram todo o seu armazenamento para armazenamento Nearline e Coldline e, em seguida, os custos aumentaram. Eles rotineiramente acessam a maioria dos objetos em qualquer dia dado. Qual é uma possível razão para os custos de armazenamento serem maiores do que o esperado?

- A. Nearline e Coldline incorrem em cobranças de acesso.
- B. Cobranças de transferência estão envolvidas.
- C. Cobranças de saída estão envolvidas.
- D. Nenhuma das opções acima.

Capítulo 12

Implantando Armazenamento no Google Cloud

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 3.4 Implantando e implementando soluções de dados
- ✓✓ 4.4 Gerenciando soluções de armazenamento e banco de dados

Neste capítulo, discutiremos como criar sistemas de armazenamento de dados em vários produtos do Google Cloud, incluindo Cloud SQL, Cloud Datastore, BigQuery, Bigtable, Cloud Spanner, Cloud Pub/Sub, Cloud Dataproc e Cloud Storage. Você aprenderá a criar bancos de dados, buckets e outras estruturas básicas de dados, bem como a realizar tarefas de gerenciamento chave, como fazer backup de dados e verificar o status de trabalhos.

Implantando e Gerenciando Cloud SQL

O Cloud SQL é um serviço gerenciado de banco de dados relacional. Nesta seção, você aprenderá a fazer o seguinte:

- Criar uma instância de banco de dados.
- Conectar à instância.
- Criar um banco de dados.
- Carregar dados no banco de dados.
- Consultar o banco de dados.
- Fazer backup do banco de dados.

Usaremos uma instância MySQL nesta seção, mas os procedimentos a seguir são semelhantes para PostgreSQL e SQL Server.

Criando e Conectando a uma Instância MySQL

Descrevemos como criar e configurar uma instância MySQL no Capítulo 11, "Planejando Armazenamento na Nuvem", mas revisaremos os passos aqui.

A partir do console, navegue até o Cloud SQL e clique em Criar Instância. Escolha MySQL para abrir a página mostrada na Figura 12.1.

Após alguns minutos, a instância é criada; a lista do MySQL terá uma aparência semelhante à Figura 12.2.

FIGURE 12.1 Creating a MySQL instance

The screenshot shows the 'Create a MySQL instance' page. It includes fields for Instance ID, Password, Database version (MySQL 8.0), and a summary table with instance details like Region, DB Version, vCPUs, Memory, Storage, Network throughput, Disk throughput, IOPS, Connections, Backup, Availability, and Point-in-time recovery.

Region	us-central1 (Iowa)
DB Version	MySQL 8.0
vCPUs	4 vCPU
Memory	26 GB
Storage	100 GB
Network throughput (MB/s)	1,000 of 2,000
Disk throughput (MB/s)	Read: 48.0 of 240.0 Write: 48.0 of 240.0
IOPS	Read: 3,000 of 15,000 Write: 3,000 of 15,000
Connections	Public IP
Backup	Automated
Availability	Multiple zones (Highly available)
Point-in-time recovery	Enabled

Instance info

Instance ID: (Required)

Password:

No password:

Database version: MySQL 8.0

Choose region and zonal availability

For better performance, keep your data close to the services that need it. Region is permanent, while zone can be changed any time.

Region: us-central1 (Iowa)

Zonal availability:

- Single zone: In case of outage, no failover. Not recommended for production.
- Multiple zones (Highly available): Automatic failover to another zone within your selected region. Recommended for production instances. Increases cost.

SPECIFY ZONES

Customize your instance

You can also customize instance configurations later

SHOW CONFIGURATION OPTIONS

CREATE INSTANCE **CANCEL**

FIGURE 12.2 A listing of MySQL instances

The screenshot shows a table of MySQL instances. It includes columns for Instance ID, Type, Public IP address, Private IP address, Instance connection name, and Actions. One instance, 'ace-exam-mysql', is listed with MySQL 8.0 type and IP 35.238.89.86.

Instance ID	Type	Public IP address	Private IP address	Instance connection name	Actions
ace-exam-mysql	MySQL 8.0	35.238.89.86		scenic-energy-33502...	

Depois que o banco de dados é criado, você pode se conectar iniciando o Cloud Shell e usando o comando `gcloud sql connect`. Este comando requer o nome da instância a qual conectar e, opcionalmente, um nome de usuário e senha. É uma boa prática não especificar uma senha na linha de comando. Em vez disso, você será solicitado a inseri-la, e ela não será exibida enquanto você digita. Você pode ver uma mensagem sobre permitir a listagem do seu endereço IP; isso é uma medida de segurança e permitirá que você se conecte à instância a partir do Cloud Shell.

Para se conectar à instância chamada ace-exam-mysql, use o seguinte comando:

```
gcloud sql connect ace-exam-mysql --user=root
```

Isso abre um prompt de linha de comando para a instância MySQL.

FIGURE 12.3 Command-line prompt to work with MySQL after connecting using gcloud sql connect

```
dan@cloudshell:~ (scenic-energy-335022)$ gcloud sql connect ace-exam-mysql --user=root
Allowlisting your IP for incoming connection for 5 minutes...done.
Connecting to database with SQL user [root].Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 257320
Server version: 8.0.26-google (Google)

Copyright (c) 2000, 2022, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> ■
```

Criando um Banco de Dados, Carregando Dados e Consultando Dados

No ambiente de linha de comando do MySQL, você usa comandos do MySQL, não comandos do gcloud. O MySQL usa SQL padrão, então o comando para criar um banco de dados é CREATE DATABASE. Você indica o banco de dados com o qual trabalhar (pode haver muitos em uma única instância) usando o comando USE. Por exemplo, para criar um banco de dados e defini-lo como o banco de dados padrão com o qual trabalhar, use isto:

```
CREATE DATABASE ace_exam_book;
```

```
USE ace_exam_book
```

Você pode então criar uma tabela usando CREATE TABLE. Os dados são inseridos usando o comando INSERT. Por exemplo, os seguintes comandos criam uma tabela chamada books e inserem duas linhas:

```
CREATE TABLE books (title VARCHAR(255), num_chapters INT, entity_id
INT NOT NULL AUTO_INCREMENT, PRIMARY KEY (entity_id));
```

```
INSERT INTO books (title,num_chapters) VALUES ('ACE Exam Study Guide',
18);
```

```
INSERT INTO books (title,num_chapters) VALUES ('Architecture Exam Study
Guide', 18);
```

Para consultar a tabela, você usa o comando SELECT. Aqui está um exemplo:

```
SELECT * FROM books;
```

Este comando listará todas as linhas na tabela.

FIGURE 12.4 Listing the contents of a table in MySQL

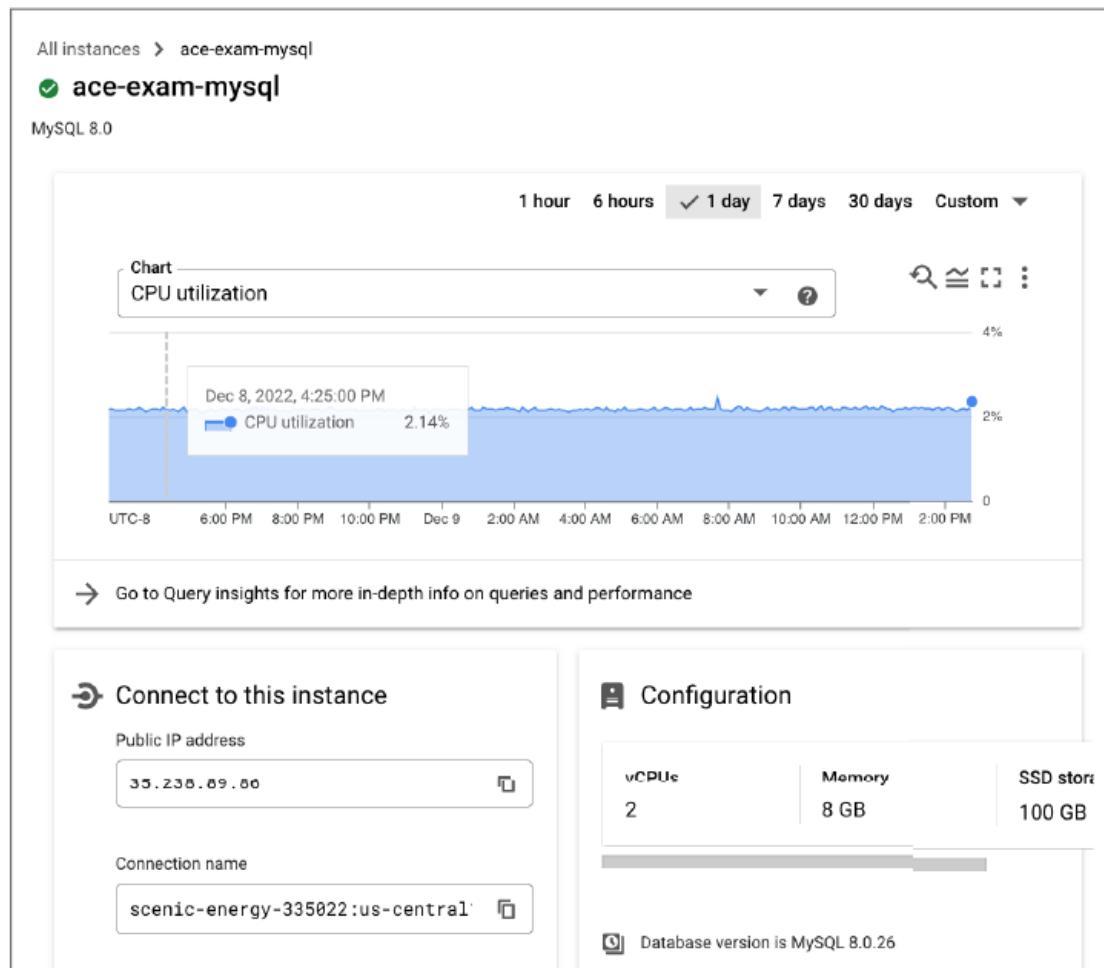
```
mysql> SELECT * FROM books;
+-----+-----+-----+
| title          | num_chapters | entity_id |
+-----+-----+-----+
| ACE Exam Study Guide |      18 |      1 |
| Architecture Exam Study Guide |      18 |      2 |
+-----+-----+-----+
2 rows in set (0.00 sec)
```

Fazendo Backup do MySQL no Cloud SQL

O Cloud SQL permite backups sob demanda e automáticos.

Para criar um backup sob demanda, clique no nome da instância na página Instâncias no console. Isso exibirá a página de Detalhes da Instância.

FIGURE 12.5 Partial listing of MySQL Instance Details page with vertical menu displayed on the left



Clique na opção de menu Backups para exibir a página Backups.

FIGURE 12.6 Create Backup button

The screenshot shows the AWS SQL interface. On the left, there's a sidebar with options: Overview, Connections, Users, Databases, and Backup (which is selected). The main area shows the primary instance 'ace-exam-mysql' (MySQL 8.0). It has a 'Settings' section with various backup configurations like 'Automated backups' (Enabled), 'Backups window' (12:00 PM – 4:00 PM (UTC-7)), and 'Days of logs retained' (7). At the bottom, there's a 'CREATE BACKUP' button and a table showing a single backup entry: 'Jun 24, 2022, 11:53:01 AM' (On-demand, Region: us-west1).

Clicar em Criar Backup abre a janela mostrada.

FIGURE 12.7 Assign a description to a backup and create it.

The dialog box is titled 'Create a Backup'. It contains a note: 'You'll be billed \$0.08/GB per month for data storage.' Below is a text input field labeled 'Describe this backup (optional)' with a character count of '0 / 140'. A note below says 'You can make a note here to help you identify this backup.' At the bottom are 'LOCATION OPTIONS' dropdown, 'CREATE' (highlighted in blue), and 'CANCEL' buttons.

Preencha a descrição opcional e clique em Criar. Quando o backup estiver completo, ele aparecerá na lista de backups.

FIGURE 12.8 Listing of backups available for this instance

The screenshot shows the 'Settings' page for a Cloud SQL instance. At the top, it displays automated backup settings: 'Automated backups' is 'Enabled', 'Backups window' is '12:00 PM - 4:00 PM (UTC-7)', 'Automated backups retained' is '7', 'Point-in-time recovery' is 'Enabled', 'Days of logs retained' is '7', and 'Location' is 'Multi-region: us'. Below these settings is a blue 'CREATE BACKUP' button. Underneath the button is a 'Filter backups' section with a dropdown menu set to 'Created'. A table lists two backups:

Created	Type	Location	Description
Jun 24, 2022, 12:20:43 PM	On-demand	MultiRegion: us	Initial backup
Jun 24, 2022, 11:53:01 AM	On-demand	Region: us-west1	Taking a backup after instance creation

Você também pode criar um backup usando o comando `gcloud sql backups`, que tem esta forma:

```
gcloud sql backups create --async --instance [NOME_DA_INSTÂNCIA]
```

Aqui, `[NOME_DA_INSTÂNCIA]` é o nome, como `ace-exam-mysql`, e o parâmetro `--async` é opcional.

Para criar um backup sob demanda para a instância `ace-exam-mysql`, use o seguinte comando:

```
gcloud sql backups create --async --instance ace-exam-mysql
```

Você também pode ter backups automáticos criados pelo Cloud SQL.

A partir do console, navegue até a página da Instância Cloud SQL, clique no nome da instância e depois clique em Editar Instância. Abra a seção de Backups Automáticos Habilitados e preencha os detalhes de quando criar os backups. Você deve especificar um intervalo de tempo para quando os backups automáticos devem ocorrer. Você também pode habilitar o registro binário, que é necessário para recursos mais avançados, como recuperação ponto-a-ponto.

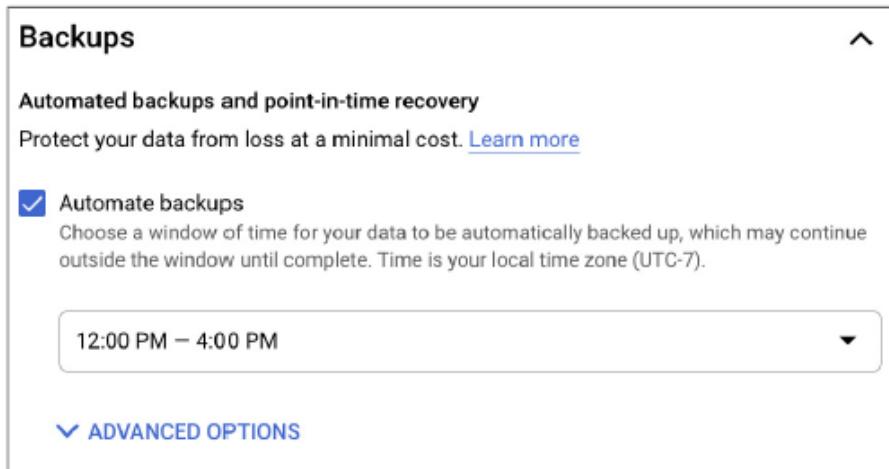
Para habilitar backups automáticos a partir da linha de comando, use o comando `gcloud`:

```
gcloud sql instances patch [NOME_DA_INSTÂNCIA] --backup-start-time [HH:MM]
```

Para esta instância de exemplo, você poderia executar backups automáticos às 1:00 da manhã com o seguinte comando:

```
gcloud sql instances patch ace-exam-mysql --backup-start-time 01:00
```

FIGURE 12.9 Enabling automatic backups in Cloud Console



Implantando e Gerenciando Firestore

O Capítulo 11 descreveu como inicializar um banco de dados de documentos Firestore. Agora, você verá como criar entidades e adicionar propriedades a um banco de dados de documentos. Você também revisará operações de backup e restauração. O Cloud Firestore é a última geração do Cloud Datastore. O Cloud Firestore possui dois modos: Nativo e Modo Datastore.

Os recursos do Cloud Firestore incluem consistência forte, modelo de dados de documento, atualizações em tempo real e bibliotecas de clientes móveis e web. Atualizações em tempo real e bibliotecas de clientes móveis e web estão disponíveis apenas no modo nativo. O modo Datastore pode escalar para milhões de escritas por segundo e é uma boa opção para um armazenamento de dados de documentos quando você não precisa dos recursos em tempo real ou móveis do modo Nativo. O modo Datastore também suporta a linguagem de consulta GQL, que é semelhante ao SQL.

Adicionando Dados a um Banco de Dados Firestore

Você adiciona dados a um banco de dados Firestore no Modo Nativo usando a opção Iniciar Coleção na seção Firestore do console. A estrutura de dados Coleções é análoga a um esquema em bancos de dados relacionais.

Você cria uma entidade clicando em Iniciar Coleção e preenchendo o formulário que aparece. Aqui, você fornecerá um ID de coleção e, em seguida, adicionará documentos, que são pares chave-valor com um tipo de dado no valor.

Após criar entidades, você pode visualizar dados no console.

FIGURE 12.10 Adding data to a Firestore collection

Start a collection

A collection is a set of one or more documents that contain data. Start a collection at this path by adding its first document. [Learn more](#)

Give the collection an ID

Parent path /

Collection ID *
ace_exam_questions

Choose an ID that describes the documents you'll add to this collection.

Add its first document [?](#)

Document ID YncGFFXRCua4TU0brrHB
Assigned automatically. Customize if needed.

Field name exam_chapter	Field type string	Field value 12
Field name title	Field type string	Field value "Managing Data"

+ ADD FIELD

SAVE **SAVE & ADD ANOTHER** **CANCEL**

FIGURE 12.11 Viewing data in Firestore, Native mode

Data			Cloud Firestore in Native mode	Database location: us-west1
/ > ace_exam_questions > YncGFFXRCua4TU0brrHB				
Root	ace_exam_questions	VneGFFXRCua4TU0brrHB		
+ START COLLECTION	+ ADD DOCUMENT	+ START COLLECTION		
ace_exam_questions	30mJlwGNizYuMwFAnAjq	exam_chapter: "12"		
	YncGFFXRCua4TU0brrHB	title: "Managing Data"		

Fazendo Backup do Firestore

Para fazer backup de um banco de dados Firestore, você precisa criar um bucket do Cloud Storage para conter um arquivo de backup e conceder permissões apropriadas aos usuários que realizam o backup.

Você pode criar um bucket para backups usando o comando gsutil:

```
gsutil mb gs://[NOME_DO_BUCKET]/
```

, [NOME DO BUCKET] é o nome, como ace_exam_backups. Em nosso exemplo, usamos ace_exam_backups e criamos esse bucket usando o seguinte:

```
gsutil mb gs://ace_exam_backups/
```

Usuários que criam backups precisam da permissão datastore.databases.export. (O Cloud Datastore foi renomeado para Cloud Firestore, mas, no momento da escrita, as funções do IAM ainda se referem ao Datastore.) Se você está importando dados, precisará de datastore.databases.import.

A função de administrador de importação/exportação do Cloud Datastore possui ambas as permissões; veja o Capítulo 17, “Configurando Acesso e Segurança”, para detalhes sobre a atribuição de funções a usuários.

Para criar um backup exportando do Firestore, você pode usar um comando como este:

```
gcloud firestore export gs://ace_exam_backups
```

Para importar um arquivo de backup, use o comando gcloud firestore import:

```
gcloud firestore import gs://ace_exam_backups
```

Implantando e Gerenciando BigQuery

O BigQuery é um serviço de banco de dados totalmente gerenciado, então o Google cuida de backups e outras tarefas administrativas básicas. Como um Engenheiro de Cloud, você ainda tem algumas tarefas administrativas ao trabalhar com o BigQuery. Duas dessas tarefas são estimar o custo de uma consulta e verificar o status de um trabalho.

FIGURE 12.12 The BigQuery console

The screenshot shows the Google Cloud BigQuery console interface. On the left, the Explorer sidebar lists projects and datasets, with 'nyc_citi_bike_trips' selected. Inside it, 'citibike_stations' is highlighted. The main area displays the schema for the 'citibike_stations' table. The schema table has columns for Field name, Type, Mode, Collation, Policy Tag, and Description. The fields listed are station_id (INTEGER, REQUIRED), name (STRING, NULLABLE), short_name (STRING, NULLABLE), latitude (FLOAT, NULLABLE), and longitude (FLOAT, NULLABLE). The descriptions provide details about each field's purpose and constraints.

Field name	Type	Mode	Collation	Policy Tag	Description
station_id	INTEGER	REQUIRED			Unique identifier of a station.
name	STRING	NULLABLE			Public name of the station.
short_name	STRING	NULLABLE			Short name or other type of identifier, as used by the data publisher.
latitude	FLOAT	NULLABLE			The latitude of station. The field value must be a valid WGS 84 latitude format.
longitude	FLOAT	NULLABLE			The longitude of station. The field value must be a valid WGS 84 longitude format.

Estimando o Custo de Consultas no BigQuery

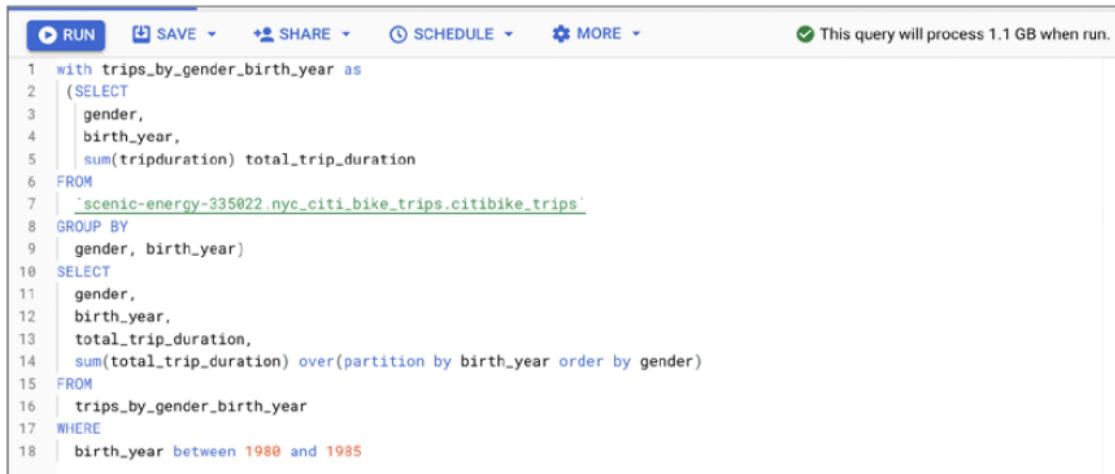
No console, escolha BigQuery no menu de navegação principal para exibir a interface de consulta do BigQuery, como mostrado parcialmente na Figura 12.12.

Aqui você pode inserir uma consulta no Editor de Consultas, como uma consulta sobre nomes e gêneros na tabela usa_1910_2013, conforme mostrado na Figura 12.13.

Observe no canto superior direito que o BigQuery fornece uma estimativa de quanto dados serão escaneados. Você também pode usar a linha de comando para obter essa estimativa usando o comando bq com a opção --dry-run:

```
bq --location=[LOCALIZAÇÃO] query --use_legacy_sql=false --dry_run [SQL_QUERY]
```

FIGURE 12.13 Example query with estimated amount of data scanned



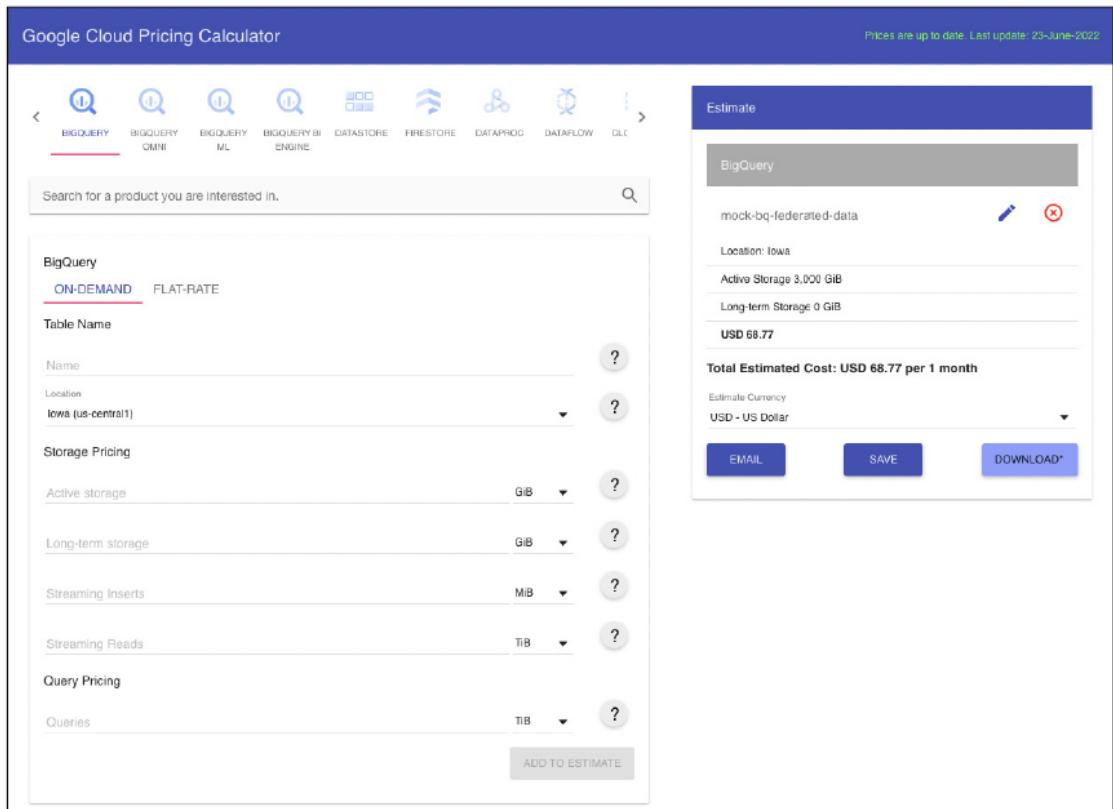
The screenshot shows the BigQuery SQL editor interface. At the top, there are buttons for RUN, SAVE, SHARE, SCHEDULE, and MORE. A status message indicates "This query will process 1.1 GB when run." The main area contains the following SQL code:

```
1 with trips_by_gender_birth_year as
2   (SELECT
3     gender,
4     birth_year,
5     sum(tripduration) total_trip_duration
6   FROM
7     `scenic-energy-335022.nyc_citi_bike_trips.citibike_trips`
8   GROUP BY
9     gender, birth_year)
10  SELECT
11    gender,
12    birth_year,
13    total_trip_duration,
14    sum(total_trip_duration) over(partition by birth_year order by gender)
15  FROM
16    trips_by_gender_birth_year
17  WHERE
18    birth_year between 1980 and 1985
```

Aqui, [LOCALIZAÇÃO] é o local em que você criou o conjunto de dados que está consultando, e [SQL_QUERY] é a consulta SQL que você está estimando.

Você pode usar esse número com a Calculadora de Preços para estimar o custo. A Calculadora de Preços está disponível em <https://cloud.google.com/products/calculator>. Após selecionar BigQuery, navegue até a aba On-Demand, insira o nome da tabela que você está consultando, defina a quantidade de armazenamento para 0 e, em seguida, insira o tamanho da consulta na linha de Consultas da seção de Precificação de Consultas. Certifique-se de usar a mesma unidade de tamanho exibida no console do BigQuery. Quando você clicar em Adicionar à Estimativa, a Calculadora de Preços exibirá o custo (veja a Figura 12.14).

FIGURE 12.14 Using the Pricing Calculator to estimate the cost of a query



Visualizando Trabalhos no BigQuery

Trabalhos no BigQuery são processos usados para carregar, exportar, copiar e consultar dados. Trabalhos são automaticamente criados quando você inicia qualquer uma dessas operações.

Para visualizar o status dos trabalhos, navegue até o console do BigQuery e clique em Histórico Pessoal ou Histórico do Projeto na seção inferior da janela de edição. Observe na Figura 12.15 que o trabalho no topo da lista tem uma marca de verificação, indicando que o trabalho foi concluído com sucesso. Isso é um exemplo de uma visualização expandida de uma entrada de trabalho. Abaixo disso está um resumo em linha única de um trabalho que falhou. A falha é indicada pelo ícone de ponto de exclamação ao lado do ID do trabalho.

FIGURE 12.15 A listing of job statuses in BigQuery

PERSONAL HISTORY		PROJECT HISTORY		SAVED QUERIES		REFRESH	
Filter Enter property name or value							
Job ID		Creation time		Owner		Type	Summary
✓ bquxjob_4428708e_1818e183371		Jun 22, 2022, 6:06:00 PM		dan@sullivanlearninggroup...		QUERY	SELECT s.station_id, count(*) FROM `scenic-energy-3350...`
✓ bquxjob_3fce001_1818e12b794		Jun 22, 2022, 6:00:00 PM		dan@sullivanlearninggroup...		QUERY	SELECT * FROM `scenic-energy-335022.nyc_citi_bike_tr...`
✓ bquxjob_48ff7fcf_18182d33afb		Jun 20, 2022, 1:34:50 PM		dan@sullivanlearninggroup...		QUERY	SELECT * FROM `scenic-energy-335022.google_analytic...`
✗ bquxjob_d929b4d_1817e655ba0		Jun 19, 2022, 6:23:44 PM		dan@sullivanlearninggroup...		QUERY	SELECT * FROM Unnest((SELECT (1,foo),(2,bar),(3,b...))

Você também poderia visualizar o status de um trabalho no BigQuery usando o comando `bq show`. Por exemplo, o seguinte comando mostra o status do trabalho especificado:

```
bq --location=US show -j gcpacer-project:US.bquijob_119adae7_167c373d5c3
```

Implantando e Gerenciando Cloud Spanner

Agora, vamos voltar nossa atenção para o Cloud Spanner, o banco de dados relacional global. Nesta seção, você criará um banco de dados, definirá um esquema, inserirá alguns dados e, em seguida, fará uma consulta.

Primeiro, você criará uma instância do Cloud Spanner. Navegue até a página do Cloud Spanner no console e clique em Criar Instância. Isso exibirá a página mostrada na Figura 12.16.

FIGURE 12.16 Creating a Cloud Spanner instance

The screenshot shows the 'Create an instance' page for Cloud Spanner. It has several sections:

- Name your instance:** Fields for 'Instance name' (ace-exam-spanner) and 'Instance ID' (ace-exam-spanner). A note says the name must be 4-30 characters long and can include lowercase letters, numbers, and hyphens.
- Choose a configuration:** A note about node location affecting cost, performance, and replication. It offers 'Regional' or 'Multi-region' options, with 'Regional' selected and 'us-west1 (Oregon)' chosen.
- Allocate compute capacity:** A note about compute capacity determining throughput, queries per second (QPS), and storage limits. It shows 'Unit' as 'Processing units' and 'Quantity' as '100'. A note says integers only, up to 1,000.
- Summary:** A table showing instance details:

Configuration	us-west1 (Oregon)
Replicas	3 read-write replicas in 3 separate zones within the region us-west1
Availability	99.99% availability SLA
Compute cost	\$0.09 per hour Save up to 20% by committing to 1 year and 40% by committing to 3 years using Committed Use Discounts. Learn more
Storage cost	\$0.30 per GB/month
Maximum storage capacity	410 GB
- Buttons:** 'CREATE' (highlighted in blue) and 'CANCEL'.

Em seguida, você precisa criar um banco de dados na instância. Clique em Criar Banco de Dados no topo da página de Detalhes da Instância para mostrar uma página semelhante à Figura 12.17.

FIGURE 12.17 Create a database within a Cloud Spanner instance.

← Create a database in ace-exam-spanner

Name your database

Enter a permanent name for your database of at least two characters, starting with a letter.

Database name * Lowercase letters, numbers, hyphens, underscores allowed

Select database dialect

Choose between Google Standard SQL and PostgreSQL dialects for your Spanner database.

Google Standard SQL
 PostgreSQL

Define your schema (optional)

Add Spanner Data Definition Language SQL statements below. Separate statements with a semicolon. [Learn more](#)

DDL TEMPLATES ▾ SHORTCUTS

Press Alt+F1 for Accessibility Options.

```
1 CREATE TABLE <table_name> (
2   | <col_name> <col_type>,
3   ) PRIMARY KEY (<col_name>);
```

▼ SHOW ENCRYPTION OPTIONS

CREATE CANCEL

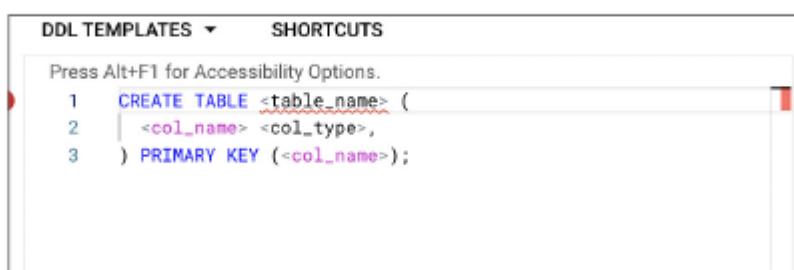
Ao criar um banco de dados, você precisará usar a linguagem de definição de dados SQL (DDL) para definir a estrutura das tabelas. SQL DDL é o conjunto de comandos SQL para criar tabelas, índices e outras estruturas de dados (veja a Tabela 12.1). A Figura 12.18 mostra um exemplo do uso de modelos de DDL fornecidos pelo Google Cloud. Neste caso, o modelo para criar uma tabela é exibido.

TABLE 12.1 SQL data definition commands

Command	Description
CREATE TABLE	Creates a table with columns and data types specified
CREATE INDEX	Creates an index on the specified column(s)
ALTER TABLE	Changes table structure
DROP TABLE	Removes the table from the database schema
DROP INDEX	Removes the index from the database schema

Além de criar uma tabela, outros modelos são mostrados na Figura 12.19.

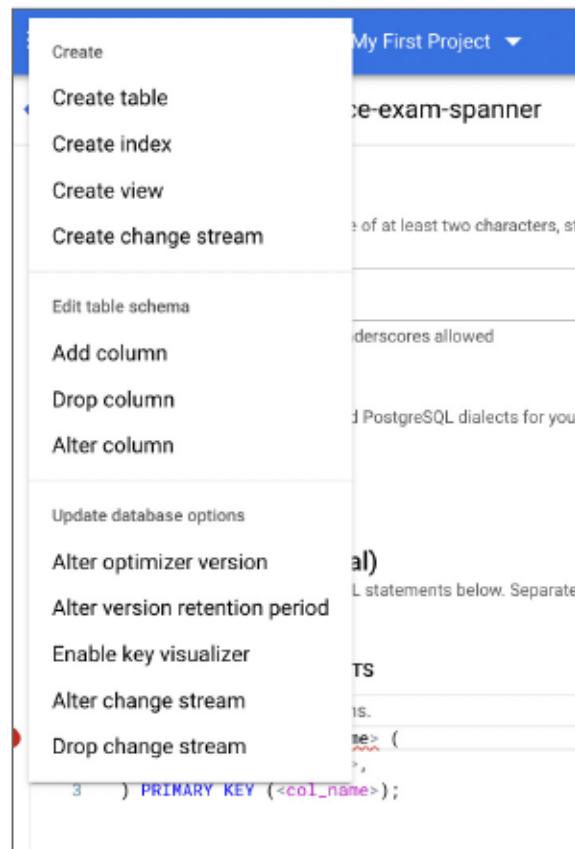
FIGURE 12.18 Creating a table using a DDL template



Uma vez que uma tabela é criada, você pode visualizar a estrutura e propriedades da tabela, conforme mostrado na Figura 12.20.

A partir da descrição do esquema da tabela, você pode navegar até o Cloud Logging para ver um histórico de mudanças na tabela, conforme mostrado na Figura 12.21.

FIGURE 12.19 DDL templates available to help you create database objects in Spanner



Finalmente, você pode revisar e adicionar papéis relacionados ao Spanner aos principais usuários do console do Spanner. A partir da lista de instâncias do Spanner, selecione a caixa de verificação para a instância. Um painel aparecerá à direita semelhante à Figura 12.22.

O Cloud Spanner é um serviço de banco de dados gerenciado, então você não terá que aplicar patches, fazer backups ou executar outras tarefas básicas de administração de dados. Suas tarefas, e aquelas dos modeladores de dados e engenheiros de software, se concentrarão no design de tabelas e consultas.

FIGURE 12.20 Details of the table created in Spanner

The screenshot shows the Google Cloud Platform Spanner Schema page for a table named `my_table`. The top navigation bar shows the path: All instances > INSTANCE ace-exam-spanner: Overview > GOOGLE STANDARD SQL DATABASE ace-spanner-db1: Overview > TABLE my_table: Schema.

Schema

Name	my_table				
Schema updates	<input checked="" type="checkbox"/> Update completed To view all updates go to Cloud Logging .				
Primary Key(s): <code>id_column (asc)</code>					
Or	Column	Type	Nullable	Order	Watched by
Or	<code>id_column</code>	INT64	No	asc	
	<code>name_column</code>	STRING(MAX)	Yes	—	

[SHOW EQUIVALENT DDL](#)

Interleaved tables

There are no tables in my_table. Add a table to get started.

[ADD TABLE](#)

FIGURE 12.21 Log of changes to Spanner table

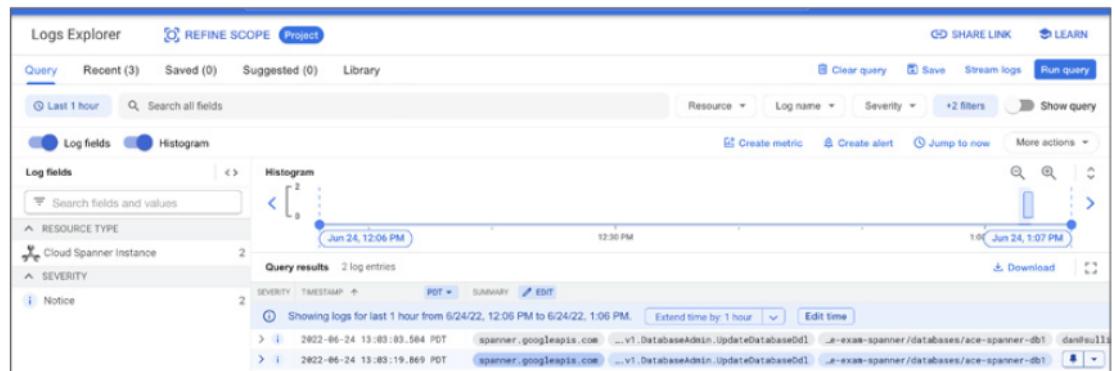


FIGURE 12.22 From the Show Info panel, you can view and manage Spanner-related roles.

The screenshot shows a 'Show Info' panel titled 'ace-spanner-db1'. At the top, there's a message: 'Edit or delete permissions below or "Add Principal" to grant new'. Below it is a button labeled '+ ADD PRINCIPAL'. A toggle switch is turned on, with the label 'Show inherited permissions'. A 'Filter' input field is present. The main area lists roles/principals under 'Role / Principal': 'Cloud Data Fusion API Service Agent (1)', 'Editor (3)', and 'Owner (1)'. To the right of the list is an 'Inheritance' link. The entire interface has a light gray background with white and blue text.

Implantando e Gerenciando Cloud Pub/Sub

Duas tarefas são necessárias para implantar uma fila de mensagens Pub/Sub: criar um tópico e criar uma assinatura. Um tópico é uma estrutura onde as aplicações podem enviar mensagens. O Pub/Sub recebe as mensagens e as mantém até que sejam lidas por uma aplicação. As aplicações leem mensagens usando uma assinatura.

O primeiro passo para trabalhar com o Pub/Sub é navegar até a página do Pub/Sub no Cloud Console. A primeira vez que você usa o Pub/Sub, o formulário será semelhante à Figura 12.23.

Depois de clicar em Criar um Tópico, você verá uma lista de assinaturas, conforme mostrado na Figura 12.24.

Você verá uma lista de tópicos exibida na página Tópicos após criar o primeiro tópico, conforme mostrado na Figura 12.25.

Para criar uma assinatura para um tópico, clique no ícone de reticências no final da linha de resumo do tópico na listagem. O menu que aparece inclui a opção Criar Assinatura (veja a Figura 12.26). Clique em Criar Assinatura para criar uma assinatura para esse tópico. Isso exibirá uma página como a mostrada na Figura 12.27.

FIGURE 12.23 Creating a Pub/Sub topic

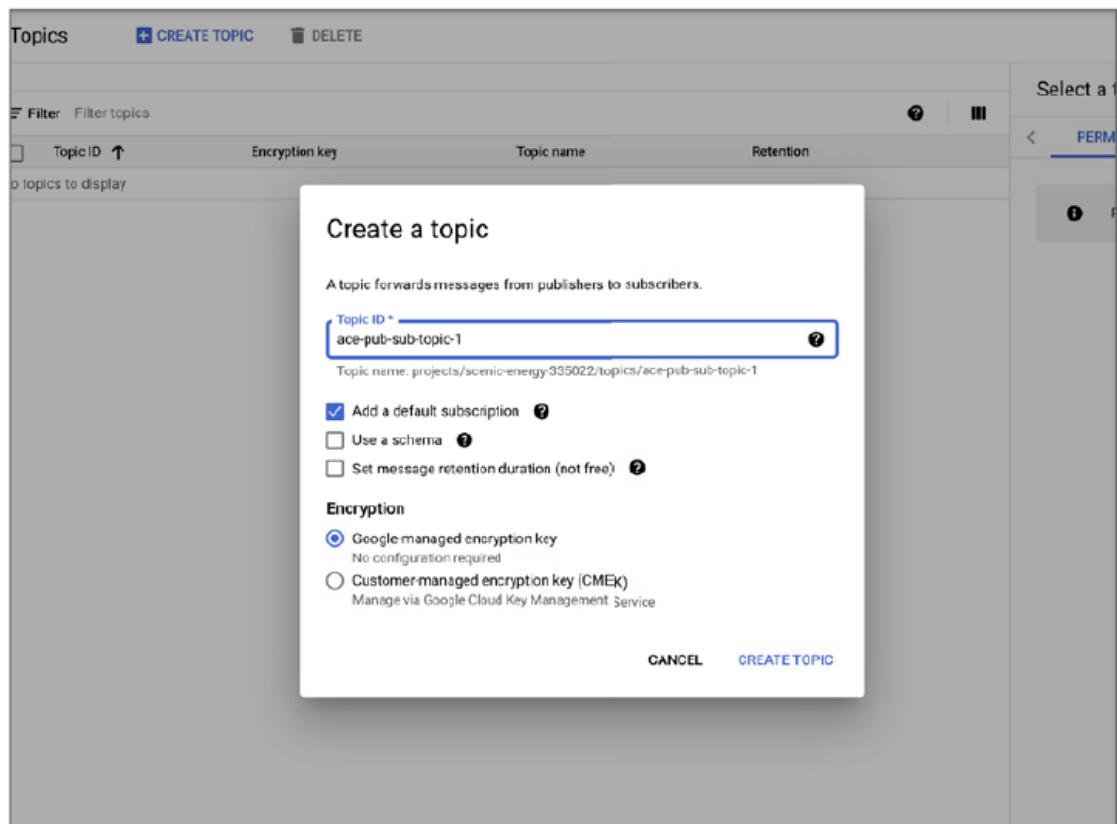


FIGURE 12.24 List of subscriptions

Subscriptions							
		Subscription ID	Delivery type	Topic name	Ack deadline	Retention	N
Filter Filter subscriptions							
<input type="checkbox"/>	State	Subscription ID	Delivery type	Topic name	Ack deadline	Retention	N
<input checked="" type="checkbox"/>	ace-pub-sub-topic-1-sub	ace-pub-sub-topic-1-sub	Pull	projects/scenic-e...	10 seconds	7 days	

Para criar uma assinatura, especifique um nome de assinatura e o tipo de entrega. As assinaturas podem ser do tipo pull, em que a aplicação lê de um tópico, ou push, em que a assinatura envia mensagens para um endpoint. Se você quiser usar uma assinatura push, precisará especificar a URL de um endpoint para receber a mensagem.

FIGURE 12.25 Subscription details

The screenshot shows the 'Subscription details' page for 'ace-pub-sub-topic-1-sub'. At the top, there are buttons for 'EDIT', 'CREATE SNAPSHOT', 'REPLAY MESSAGES', and 'PURGE MESSAGES'. Below this, the 'Subscription name' is listed as 'projects/scenic-energy-335022/subscriptions/ace-pub-sub-topic-1-sub' with a copy icon. The 'Subscription state' is 'active' with a checkmark. The 'Topic name' is 'projects/scenic-energy-335022/topics/ace-pub-sub-topic-1' with a copy icon. Below these fields are tabs for 'MESSAGES', 'METRICS', and 'DETAILS'. A note says: 'Click Pull to view messages and temporarily delay message delivery to other subscribers. Select Enable ACK messages and then click ACK next to the message to permanently prevent message delivery to other subscribers.' Under the 'MESSAGES' tab, there are buttons for 'PULL' and 'Enable ack messages'. A 'Filter' button is available to 'Filter messages'. The message list table has columns for 'Publish time', 'Attribute keys', 'Message body', 'Ordering key', and 'Ack ↑'. A message count of 'No message found yet' is displayed.

FIGURE 12.26 Creating a subscription to a topic

The screenshot shows the 'Topics' page. At the top, there are buttons for 'CREATE TOPIC' and 'DELETE'. A filter bar labeled 'Filter topics' is present. The table lists topics with columns for 'Topic ID', 'Encryption key', 'Topic name', and 'Retention'. One topic is listed: 'ace-pub-sub-topic-1' with 'Google-managed' encryption and 'projects/scenic-energy-335022/topics/ace-pub-sub-topic-1' as the topic name. To the right of the table is a 'PERMISSIONS' section with a 'Select a topic' dropdown and a context menu. The context menu includes options: 'Create subscription', 'Create snapshot', 'Import from', 'Export to', 'Trigger Cloud Function', 'Update labels', 'View permissions', 'View storage policies', and 'Delete'.

Uma vez que uma mensagem é lida, a aplicação que lê a mensagem reconhece o recebimento da mensagem. O Pub/Sub aguardará o período de tempo especificado no parâmetro Prazo de Reconhecimento. O tempo de espera pode variar de 10 a 600 segundos.

Você também pode especificar um período de retenção, que é o tempo de conservação de uma mensagem que não pode ser entregue. Após o período de retenção passar, as mensagens são deletadas do tópico.

Quando você terminar de criar uma assinatura, verá uma lista de assinaturas como a mostrada na Figura 12.28.

FIGURE 12.27 The options for creating a subscription

The screenshot shows the 'Create subscription' interface. At the top, there's a back arrow and the title 'Create subscription'. Below the title, a descriptive text states: 'A subscription directs messages on a topic to subscribers. Messages can be pushed to subscribers immediately, or subscribers can pull messages as needed.' The main configuration area includes:

- Subscription ID ***: 'ace-exam-subscription-1' (with a question mark icon).
- Select a Cloud Pub/Sub topic ***: 'projects/scenic-energy-335022/topics/ace-pub-sub-topic-1' (highlighted with a blue border).
- Delivery type**: A radio button group where 'Pull' is selected (indicated by a blue outline).
- Message retention duration**: A section showing 'Duration is from 10 minutes to 7 days' with dropdowns for Days (7), Hours (0), and Minutes (0).
- Retain acknowledged messages**: An unchecked checkbox with a question mark icon.
- Expiration period**: A section where 'Expire after this many days of inactivity (up to 365)' is selected (radio button highlighted with a blue outline). It shows a field with '31 Days'.
- Never expire**: An unchecked checkbox.
- Acknowledgement deadline**: A section with a question mark icon.

FIGURE 12.28 A list of subscriptions

The screenshot shows a table titled 'Subscriptions' with a 'CREATE SUBSCRIPTION' button and a 'DELETE' button. The table has columns: State, Subscription ID, Delivery type, Topic name, Ack deadline, Retention, and three more columns represented by ellipses. A 'Filter' row allows filtering by 'Subscription ID' and sorting by 'Delivery type'. Two rows are listed:

	Subscription ID	Delivery type	Topic name	Ack deadline	Retention	⋮
<input checked="" type="checkbox"/>	ace-exam-subscription-2	Pull	projects/scenic-e...	10 seconds	7 days	<input type="button"/> ⋮
<input checked="" type="checkbox"/>	ace-pub-sub-topic-1-sub	Pull	projects/scenic-e...	10 seconds	7 days	<input type="button"/> ⋮

Além de usar o console, você pode usar comandos gcloud para criar tópicos e assinaturas. Os comandos para criar tópicos e assinaturas são os seguintes:

```
gcloud pubsub topics create [NOME-DO-TÓPICO]
```

```
gcloud pubsub subscriptions create [NOME-DA-ASSINATURA] --topic [NOME-DO-TÓPICO]
```

Implantando e Gerenciando Cloud Bigtable

Como Engenheiro de Cloud, você pode precisar criar um cluster Bigtable, ou conjunto de servidores executando serviços Bigtable, bem como criar tabelas, adicionar dados e consultar esses dados.

Para criar uma instância Bigtable, navegue até o console Bigtable e clique em Criar Instância. Isso exibirá a página mostrada na Figura 12.29. (Veja o Capítulo 11 para detalhes adicionais sobre a criação de uma instância Bigtable.)

Uma vez que uma instância é criada, você pode visualizar um resumo dos dados de desempenho na página de Detalhes da Instância, como mostrado na Figura 12.30.

Muito do trabalho que você fará com o Bigtable é feito na linha de comando.

Para criar uma tabela, abra um navegador Cloud Shell e instale o comando cbt. Diferente de bancos de dados relacionais, o Bigtable é um banco de dados NoSQL e não usa o comando SQL. Em vez disso, o comando cbt tem subcomandos para criar tabelas, inserir dados e consultar tabelas.

TABLE 12.2 cbt commands

Command	Description
createtable	Creates a table
createfamily	Creates a column family
read	Reads and displays rows
ls	Lists tables and columns

Para configurar o cbt no Cloud Shell, insira estes comandos:

```
gcloud components update
```

```
gcloud components install cbt
```

FIGURE 12.29 Creating a Bigtable instance

A Bigtable instance is a container for your clusters. [Learn more](#)

1 Name your instance **\$468 per month (estimated)**
 That's about \$0.65 an hour with 0 GB stored.

2 Select your storage type **SHOW DETAILS**

3 Configure your first cluster

A cluster handles application requests for an instance. It contains nodes which determine your cluster's performance and storage limit.

Additional clusters can be added at any time.

Select a cluster ID
 ID is permanent

Cluster ID * ace-bigtable-1-c1

Select a location
 Choice is permanent. Determines where cluster data is stored. To reduce latency and increase throughput, store your data near the services that need it. [Learn more](#)

Region * **Zone**

Choose node scaling mode
 Nodes are compute resources that Bigtable uses to manage your data and perform maintenance tasks. Adding nodes helps a cluster handle larger workloads.

Scaling mode and configurations can be changed at any time.

Manual allocation
 Set your node count for fixed costs and compute resources.

For better instance performance, keep your CPU utilization under the recommended threshold for your [app profile routing policy](#). [Contact us](#) if you need to increase your quota.

Quantity * 1 **Nodes**

Autoscaling
 Let Bigtable automatically add and remove nodes.

SHOW ENCRYPTION OPTIONS

O Bigtable requer uma variável de ambiente chamada `instance` que deve ser definida incluindo-a em um arquivo de configuração CBT chamado `.cbtrc`, que é mantido no diretório `home`.

FIGURE 12.30 Instance details, including performance data

Por exemplo, para definir a instância para ace-exam-bigtable, insira este comando no prompt de linha de comando:

```
echo instance = ace-exam-bigtable >> ~/.cbtrc
```

Agora, os comandos cbt operarão nessa instância. Para criar uma tabela, emita um comando como este:

```
cbt createtable ace-exam-bt-table
```

O comando ls lista as tabelas. Aqui está um exemplo:

```
cbt ls
```

Isso exibirá uma lista de todas as tabelas. As tabelas contêm colunas, mas o Bigtable também tem um conceito de famílias de colunas. Para criar uma família de colunas chamada colfam1, use o seguinte comando:

```
cbt createfamily ace-exam-bt-table colfam1
```

Para definir o valor da célula com a coluna colfam1 em uma linha chamada row1, use o seguinte comando:

```
cbt set ace-exam-bt-table row1 colfam1:col1=ace-exam-value
```

Para exibir o conteúdo de uma tabela, use um comando de leitura como este:

```
cbt read ace-exam-bt-table
```

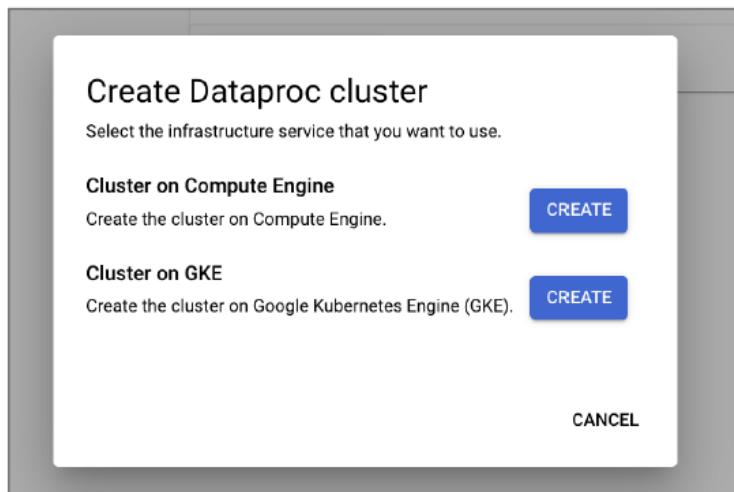
Implantando e Gerenciando Cloud Dataproc

O Cloud Dataproc é o serviço gerenciado do Google para Apache Spark e Apache Hadoop. Tanto o Spark quanto o Hadoop são projetados para aplicações de "big data". O Spark suporta análise e aprendizado de máquina, enquanto o Hadoop é adequado para aplicações de big data em lotes. Como Engenheiro de Cloud, você deve estar familiarizado com a criação de um cluster Dataproc e submissão de trabalhos para execução no cluster.

Para criar um cluster, navegue até a parte do Dataproc no Console do Cloud e selecione Criar Cluster; depois escolha a infraestrutura subjacente, que pode ser o

Compute Engine ou Google Kubernetes Engine (veja a Figura 12.31). O Google Kubernetes Engine é uma boa opção se você tem um cluster GKE existente e quer usá-lo para um cluster Spark/Hadoop gerenciado pelo Cloud Dataproc. Se você não tem um cluster GKE ou não quer executar clusters Cloud Dataproc em seus clusters GKE, então usar o Compute Engine é a opção melhor.

FIGURE 12.31 Choose an infrastructure for your cluster, either Compute Engine or Google Kubernetes Engine.



Crie um cluster Dataproc completando as opções de Criar Cluster. Você precisará especificar o nome do cluster e uma região e zona. Você também precisará especificar o modo do cluster, que pode ser Padrão, Nô Único ou Alta Disponibilidade. Nô Único é útil para desenvolvimento. Padrão tem apenas um nó mestre, então se falhar, o cluster torna-se inacessível. O modo de Alta Disponibilidade usa três mestres.

Você também precisará especificar informações de configuração da máquina para os nós mestres e os nós trabalhadores. Você especificará CPUs, memória e informações de disco. O modo do cluster determina o número de nós mestres, mas você pode escolher o número de nós trabalhadores.

Se você optar por expandir a lista de opções avançadas, você pode indicar que gostaria de usar VMs preemptivas e especificar o número de VMs preemptivas que deseja executar (não mostrado nas figuras). A Figura 12.32 mostra as opções para criar um cluster no Compute Engine, e a Figura 12.33 mostra as opções para criar um cluster no Google Kubernetes Engine.

Quando o cluster estiver em execução, você pode enviar trabalhos usando a página Enviar um Trabalho, conforme mostrado na Figura 12.34.

FIGURE 12.32 Creating a Dataproc cluster on Compute Engine

Create a Dataproc cluster on Compute Engine

- Set up cluster**
Begin by providing basic information.
- Configure nodes (optional)**
Change node compute and storage capabilities.
- Customize cluster (optional)**
Add cluster properties, features, and actions.
- Manage security (optional)**
Change access, encryption, and security settings.

Name
Cluster Name * ?

Location
Region * ? Zone * ?

Cluster type

Standard (1 master, N workers)

Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Autoscaling
Automates cluster resource management based on an autoscaling policy.

Policy ▼

Enhanced Flexibility Mode
Dataproc Enhanced Flexibility Mode (EFM) manages shuffle data to minimize job progress delays caused by the removal of nodes from a running cluster. EFM offloads shuffle data in one of two user-selectable modes, primary worker shuffle and Hadoop Compatible File System (HCFS) shuffle. [Learn more](#)

! An autoscaling policy must be selected to configure EFM.

Versioning
Use a custom image to load pre-installed packages. [Learn more](#)

CREATE **CANCEL**

EQUIVALENT COMMAND LINE ▼

Você precisará especificar o cluster no qual executar o trabalho e o tipo de trabalho, que pode ser Spark, PySpark, SparkR, Hive, Spark SQL, Pig ou Hadoop. Os arquivos JAR são os programas Java que serão executados, e a Classe Principal ou JAR é o nome da função ou método que deve ser invocado para iniciar o trabalho. Se você escolher PySpark, enviará um programa Python; se enviar SparkR, enviará um programa R. Ao executar Hive ou SparkSQL, você enviará arquivos de consulta. Você também pode passar argumentos opcionais.

FIGURE 12.33 Creating a Dataproc cluster on Google Kubernetes Engine

← Create a Dataproc cluster on GKE

- Set up cluster
Begin by providing basic information.
- Configure Node pools
Change the shape and size of your Kubernetes node pools.
- Customize cluster (optional)
Add cluster properties, features, and actions.

CREATE **CANCEL**

Name
Cluster Name * gke-cluster-f923

Location
Region * us-central1

Versioning
Image Type and Version
dataproc-2.0
Release Date
First released on 1/22/2021.
CHANGE

Kubernetes Cluster
Enter an underlying Kubernetes cluster.
Kubernetes Cluster *

Cloud Storage staging bucket
Cloud Storage staging bucket to be used for storing cluster job dependencies, job driver output, and cluster config files.
Storage staging bucket *

Workload identity
Dataproc on GKE requires the use of [GKE Workload Identity](#). If you have the necessary permissions, the 'Setup workload identity' item is enabled. Make this selection to have Google Cloud Console set up the Workload Identity bindings for you. If this selection is disabled or you decide to set up your own bindings, see [Dataproc on GKE IAM Roles and Identity](#) for more information.
 Setup workload identity

Você também pode criar modelos de fluxo de trabalho no Cloud Dataproc (veja a Figura 12.35). Modelos de fluxo de trabalho permitem definir e executar fluxos de trabalho especificados como um gráfico dirigido de trabalhos. Com modelos de fluxo de trabalho, você pode especificar se deseja usar um cluster gerenciado, o que permitiria que o fluxo de trabalho criasse um cluster, executasse os trabalhos e, em seguida, encerrasse o cluster automaticamente. Alternativamente, você pode especificar um cluster no qual executar os trabalhos. Modelos de fluxo de trabalho são úteis quando você tem que executar trabalhos complexos no Cloud Dataproc.

FIGURE 12.34 Submitting a job and choosing a job type

Job ID * job-325f9d63

Region * us-west1

Specifies the Cloud Dataproc regional service, which determines what clusters are available.

Cluster * ace-dataproc-cluster-1

Job type *

- Hadoop
- Spark
- SparkR
- PySpark
- Hive
- SparkSQL
- Pig

Archive files

Archive files are extracted in the Spark working directory. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix. Supported file types: .jar, .tar, .tar.gz, .tgz, .zip.

Arguments

Additional arguments to pass to the main class. Press Return after each argument.

Max restarts per hour

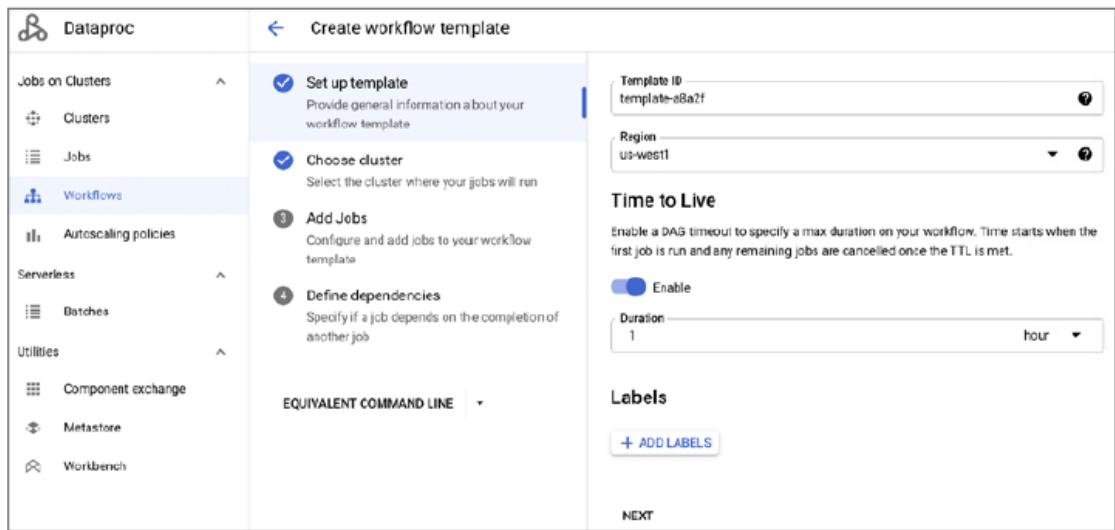
Leave blank if you don't want to allow automatic restarts on job failure. [Learn more](#)

Além de permitir a criação de clusters e fluxos de trabalho, o Cloud Dataproc suporta uma opção Spark Sem Servidor. Você pode executar trabalhos em lotes escolhendo a opção Lotes na seção Sem Servidor do painel de navegação do Dataproc, conforme mostrado na Figura 12.36. Com a opção Sem Servidor, você não precisa configurar recursos do cluster ou gerenciar clusters.

Além de usar o console, você pode criar um cluster usando o comando gcloud dataproc clusters. Aqui está um exemplo:

```
gcloud dataproc clusters create cluster-bc3d --zone us-west2-a
```

FIGURE 12.35 Creating a workflow template



Este comando criará um cluster padrão na zona us-west2-a. Você também pode especificar parâmetros adicionais para tipos de máquina, configurações de disco e outras características do cluster.

Você usa o comando gcloud dataproc jobs para enviar trabalhos da linha de comando. Aqui está um exemplo:

```
gcloud dataproc jobs submit spark --cluster cluster-bc3d --jar ace_exam_jar.jar
```

Isso enviará um trabalho executando o programa ace_exam_jar.jar no cluster cluster-bc3d.

Mundo Real

Spark para Aprendizado de Máquina

Varejistas coletam grandes volumes de dados sobre as compras dos clientes, e isso é especialmente útil para entender as preferências e interesses dos consumidores. Os sistemas de processamento de transações que coletam muito desses dados não são projetados para analisar grandes volumes de dados. Por exemplo, se os varejistas quisessem recomendar produtos aos clientes com base em seus interesses, eles poderiam construir modelos de aprendizado de máquina treinados com seus dados de vendas. O Spark possui uma biblioteca de aprendizado de máquina, chamada MLlib, que é projetada para justamente esse tipo de problema. Engenheiros podem exportar dados dos sistemas de processamento de transações, carregá-los no Spark e, em seguida, aplicar uma variedade de algoritmos de aprendizado de máquina, como agrupamento e filtragem colaborativa, para recomendações. A saída desses modelos inclui produtos que provavelmente serão de interesse de determinados clientes. São aplicações como estas que impulsionam a adoção do Spark e outras plataformas de análise.

FIGURE 12.36 Serverless options allow you to run jobs without configuring clusters.

The screenshot shows the Google Cloud Platform DataProc interface. On the left, there's a sidebar with several sections: 'Jobs on Clusters' (Clusters, Jobs, Workflows), 'Autoscaling policies', and 'Serverless' (Batches, Utilities, Component exchange, Metastore, Workbench). The 'Batches' section is currently selected. The main area is titled 'Create batch' and contains a 'Batch info' form. The 'Batch ID' field is set to 'batch-9aaf'. The 'Region' dropdown is set to 'us-central1'. Under the 'Container' section, 'Batch type' is set to 'Spark'. There are two radio button options: 'Main class' (selected) and 'Main jar URI'. Below these are sections for 'Custom container image', 'Jar files', 'Files', and 'Archive files'. At the bottom of the sidebar, there's a 'Release Notes' link.

Gerenciando Cloud Storage

No Capítulo 11, você viu como usar políticas de gerenciamento de ciclo de vida para mudar automaticamente a classe de armazenamento de um bucket. Por exemplo, você poderia criar uma política para mudar um bucket de classe de armazenamento regional para um bucket nearline após 90 dias. No entanto, pode haver momentos em que você gostaria de mudar manualmente a classe de armazenamento de um bucket. Nesses casos, você pode usar o comando gsutil rewrite e especificar a bandeira -s. Aqui está um exemplo:

```
gsutil rewrite -s[CLASSE_DE_ARMAZENAMENTO]gs://[CAMILHO_PARA_OBJETO]
```

Aqui, [CLASSE_DE_ARMAZENAMENTO] é a nova classe de armazenamento.

Outra tarefa comum com o Cloud Storage é mover objetos entre buckets. Você pode fazer isso usando o comando gsutil mv. A forma do comando é a seguinte:

```
gsutil mv gs://[NOME_DO_BUCKET_ORIGEM]/[NOME_DO_OBJETO_ORIGEM] \
gs://[NOME_DO_BUCKET_DESTINO]/[NOME_DO_OBJETO_DESTINO]
```

Aqui, [NOME_DO_BUCKET_ORIGEM] e [NOME_DO_OBJETO_ORIGEM] são o nome original do bucket e do arquivo, e [NOME_DO_BUCKET_DESTINO] e [NOME_DO_OBJETO_DESTINO] são o bucket de destino e o nome do arquivo, respectivamente.

O comando de mover também pode ser usado para renomear um objeto, similar ao comando mv no Linux. Para um objeto no Cloud Storage, você pode usar este comando:

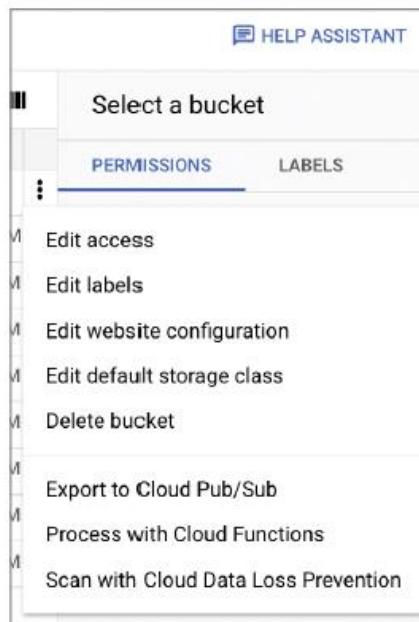
```
gsutil mv gs://[NOME_DO_BUCKET]/[NOME_ANTIGO_DO_OBJETO] \
gs://[NOME_DO_BUCKET]/[NOME_NOVO_DO_OBJETO]
```

Você também pode usar o console para realizar uma série de operações, incluindo:

- Editar acesso
- Editar etiquetas
- Deletar um bucket
- Exportar para o Cloud Pub/Sub
- Processar com o Cloud Functions
- Escanear com o serviço Cloud Data Loss Protection

O Google Cloud adicionou um comando gcloud storage ao utilitário gcloud. Ele tem funcionalidades semelhantes ao gsutil e é geralmente mais rápido tanto para uploads quanto para downloads.

FIGURE 12.37 Operations you can perform on buckets in Cloud Storage



Resumo

Neste capítulo, você aprendeu a realizar tarefas básicas de implantação e gerenciamento para vários serviços do Google Cloud, incluindo Cloud SQL, Cloud Datastore, BigQuery, Bigtable, Cloud Spanner, Cloud Pub/Sub, Cloud Dataproc e Cloud Storage. Você viu como usar o console e ferramentas de linha de comando. Enquanto o gcloud é frequentemente usado, vários dos serviços têm suas próprias ferramentas de linha de comando. Houve alguma discussão sobre como criar estruturas de banco de dados, inserir dados e consultar esses dados nos vários serviços de banco de dados. Também discutimos operações básicas de gerenciamento do Cloud Storage, como mover e renomear objetos.

Essencial para o Exame

Entenda como inicializar o Cloud SQL e o Cloud Spanner. O Cloud SQL e o Cloud Spanner são os dois bancos de dados relacionais gerenciados para sistemas de processamento de transações. O BigQuery é um banco de dados analítico projetado para data warehouse e análise. Entenda a necessidade de criar bancos de dados e tabelas. Saiba que o SQL é usado para consultar esses bancos de dados.

Entenda como inicializar o Cloud Firestore e o Cloud Bigtable. Estes são duas ofertas NoSQL. Você pode adicionar pequenas quantidades de dados ao Cloud Firestore através do console e consultá-lo com uma linguagem semelhante ao SQL chamada GQL. O Cloud Bigtable é um banco de dados de colunas largas que não suporta SQL. O Bigtable é gerenciado com a ferramenta de linha de comando cbt.

Saiba como exportar dados do BigQuery, estimar o custo de uma consulta e monitorar trabalhos no BigQuery. O BigQuery é projetado para trabalhar com data

warehouses na escala de petabytes. SQL é usado para consultar dados. Saiba como exportar dados usando o console. Entenda que o comando bq na linha de comando, não o gcloud, é a ferramenta para trabalhar com o BigQuery pela linha de comando.

Saiba como converter classes de armazenamento de buckets do Cloud Storage. Políticas de ciclo de vida podem mudar as classes de armazenamento de buckets quando eventos ocorrem, como a passagem de um período de tempo. Saiba que o gsutil rewrite é usado para mudar a classe de armazenamento de um bucket interativamente. Saiba como usar o console e a linha de comando para mover e renomear objetos.

Entenda que o Pub/Sub é uma fila de mensagens. Aplicações escrevem dados em tópicos, e aplicações recebem mensagens através de assinaturas em tópicos. Assinaturas podem ser push ou pull. Mensagens não lidas têm um período de retenção após o qual são deletadas.

Entenda que o Cloud Dataproc é um serviço gerenciado de Spark e Hadoop. Estas plataformas são usadas para análise de big data, aprendizado de máquina e trabalhos em lote de grande escala, como operações de extração, transformação e carga de grande volume. O Spark é uma boa opção para analisar dados de transação, mas os dados devem ser carregados no Spark a partir de seu sistema de origem.

Conheça as quatro ferramentas de linha de comando: gcloud, gsutil, bq e cbt. gcloud é usado para a maioria dos produtos, mas não todos. gsutil e os comandos gcloud storage mais recentes são usados para trabalhar com o Cloud Storage a partir da linha de comando. Se você quer trabalhar com o BigQuery a partir da linha de comando, precisa usar o bq. Para trabalhar com o Bigtable, você usa o comando cbt.

Questões

Você pode encontrar as respostas no Apêndice.

1. O Cloud SQL é um serviço de banco de dados relacional totalmente gerenciado, mas os administradores de banco de dados ainda precisam realizar algumas tarefas. Quais das seguintes tarefas os usuários do Cloud SQL precisam executar?
 - A. Aplicar patches de segurança
 - B. Realizar backups programados regularmente
 - C. Criar bancos de dados
 - D. Ajustar o sistema operacional para otimizar o desempenho do Cloud SQL
2. Qual dos seguintes comandos é usado para criar um backup de um banco de dados Cloud SQL?
 - A. gcloud sql backups create
 - B. gsutil sql backups create
 - C. gcloud sql create backups
 - D. gcloud sql backups export
3. Qual dos seguintes comandos executará um backup automático às 3:00 da manhã em uma instância chamada ace-exam-mysql?
 - A. gcloud sql instances patch ace-exam-mysql --backup-start-time 03:00
 - B. gcloud sql databases patch ace-exam-mysql --backup-start-time 03:00
 - C. cbt sql instances patch ace-exam-mysql --backup-start-time 03:00
 - D. bq gcloud sql instances patch ace-exam-mysql --backup-start-time 03:00
4. Qual é a linguagem de consulta usada pelo Firestore no modo Datastore?
 - A. SQL
 - B. MDX
 - C. GQL
 - D. DataFrames
5. Qual é a estrutura de linha de comando correta para exportar dados do Firestore?
 - A. gcloud firestore export collection gs://[NOME_DO_BUCKET]
 - B. gcloud firestore dump collection gs://[NOME_DO_BUCKET]
 - C. gcloud firestore export gs://[NOME_DO_BUCKET]
 - D. gcloud firestore dump gs://[NOME_DO_BUCKET]

6. Quando você insere uma consulta no formulário de consulta do BigQuery, o BigQuery analisa a consulta e exibe uma estimativa de qual métrica?

 - A. Tempo necessário para inserir a consulta
 - B. Custo da consulta
 - C. Quantidade de dados escaneados
 - D. Número de bytes transferidos entre servidores no cluster do BigQuery
7. Você quer obter uma estimativa do volume de dados escaneados pelo BigQuery a partir da linha de comando. Qual opção mostra a estrutura de comando que você deve usar?

 - A. gcloud BigQuery query estimate [SQL_QUERY]
 - B. bq --location=[LOCALIZAÇÃO] query --use_legacy_sql=false --dry_run [SQL_QUERY]
 - C. gsutil --location=[LOCALIZAÇÃO] query --use_legacy_sql=false --dry_run [SQL_QUERY]
 - D. cbt BigQuery query estimate [SQL_QUERY]
8. Você está usando o Cloud Console e quer verificar alguns trabalhos em execução no BigQuery. Você navega até a parte do BigQuery no console. Qual item de menu você clicaria para visualizar trabalhos?

 - A. Histórico Pessoal ou Histórico do Projeto.
 - B. Trabalhos Ativos.
 - C. Meus Trabalhos.
 - D. Não é possível visualizar o status do trabalho no console; você tem que usar o bq na linha de comando.
9. Você quer estimar o custo de executar uma consulta no BigQuery. Quais dois serviços dentro do Google Cloud você precisará usar?

 - A. BigQuery e Faturamento
 - B. Faturamento e Calculadora de Preços
 - C. BigQuery e Calculadora de Preços
 - D. Faturamento e comando bq
10. Você acabou de criar uma instância do Cloud Spanner. Foi-lhe atribuída a tarefa de criar uma maneira de armazenar dados sobre um catálogo de produtos. Qual é o próximo passo após criar uma instância do Cloud Spanner que você executaria para permitir carregar dados?

- A. Executar gcloud spanner update-security-patches.
 - B. Criar um banco de dados dentro da instância.
 - C. Criar tabelas para conter os dados.
 - D. Usar o console do Cloud Spanner para importar dados em tabelas criadas com a instância.
11. Você criou uma instância e banco de dados do Cloud Spanner. De acordo com as melhores práticas do Google, com que frequência você deve atualizar os pacotes da VM usando apt-get?
- A. A cada 24 horas.
 - B. A cada 7 dias.
 - C. A cada 30 dias.
 - D. Nunca; o Cloud Spanner é um serviço gerenciado.
12. Sua equipe de software está desenvolvendo uma aplicação distribuída e quer enviar mensagens de uma aplicação para outra. Uma vez que a aplicação consumidora leia uma mensagem, ela deve ser excluída. Você quer que seu sistema seja robusto a falhas, então as mensagens devem estar disponíveis por pelo menos três dias antes de serem descartadas. Qual serviço do Google Cloud é melhor projetado para suportar esse caso de uso?
- A. Bigtable
 - B. Dataproc
 - C. Cloud Pub/Sub
 - D. Cloud Spanner
13. Seu gerente pede que você configure um sistema Pub/Sub básico como uma área de teste para novos desenvolvedores aprenderem sobre sistemas de mensagens. Quais são os dois recursos dentro do Pub/Sub que você precisará criar?
- A. Tópicos e tabelas
 - B. Tópicos e bancos de dados
 - C. Tópicos e assinaturas
 - D. Tabelas e assinaturas
14. Sua empresa está lançando um serviço de IoT e receberá grandes volumes de dados de streaming. Você precisa armazenar esses dados no Bigtable. Você quer explorar o ambiente do Bigtable a partir da linha de comando. Qual comando você executaria para garantir que tem as ferramentas de linha de comando instaladas?
- A. apt-get install bigtable-tools
 - B. apt-get install cbt

- C. gcloud components install cbt
- D. gcloud components install bigtable-tools
15. Você precisa criar uma tabela chamada iot-ingest-data no Bigtable. Qual comando você usaria?
- A. cbt createtable iot-ingest-data
- B. gcloud bigtable tables create iot-ingest-data
- C. gcloud bigtable create tables iot-ingest-data
- D. gcloud create iot-ingest-data
16. O Cloud Dataproc é um serviço gerenciado para quais duas plataformas de big data?
- A. Spark e Cassandra
- B. Spark e Hadoop
- C. Hadoop e Cassandra
- D. Spark e TensorFlow
17. Seu departamento foi solicitado a analisar grandes lotes de dados todas as noites. Os trabalhos serão executados por cerca de três a quatro horas. Você quer desligar os recursos assim que a análise for concluída, então decide escrever um script para criar um cluster Dataproc todas as noites à meia-noite. Qual comando você usaria para criar um cluster chamado spark-nightly-analysis na zona us-west2-a?
- A. bq dataproc clusters create spark-nightly-analysis --zone us-west2-a
- B. gcloud dataproc clusters create spark-nightly-analysis --zone us-west2-a
- C. gcloud dataproc clusters spark-nightly-analysis --zone us-west2-a
- D. Nenhuma das opções acima
18. Você tem vários buckets contendo dados antigos que raramente são usados. Você não quer deletá-los, mas quer minimizar o custo de armazená-los. Você decide mudar a classe de armazenamento para Coldline para cada um desses buckets. Qual é a estrutura de comando que você usaria?
- A. gcloud rewrite -s [CLASSE_DE_ARMAZENAMENTO] gs://[CAMINHO_PARA_OBJETO]

B. gsutil rewrite -s [CLASSE_DE_ARMAZENAMENTO]
 gs://[CAMILHO_PARA_OBJETO]

C. cbt rewrite -s [CLASSE_DE_ARMAZENAMENTO]
 gs://[CAMILHO_PARA_OBJETO]

D. bq rewrite -s [CLASSE_DE_ARMAZENAMENTO]
 gs://[CAMILHO_PARA_OBJETO]

19. Você quer renomear um objeto armazenado em um bucket. Qual estrutura de comando você usaria?

- A. gsutil cp gs://[NOME_DO_BUCKET]/[NOME_ANTIGO_DO_OBJETO]
 gs://[NOME_DO_BUCKET]/[NOVO_NOME_DO_OBJETO]
- B. gsutil mv gs://[NOME_DO_BUCKET]/[NOME_ANTIGO_DO_OBJETO]
 gs://[NOME_DO_BUCKET]/[NOVO_NOME_DO_OBJETO]
- C. gsutil mv gs://[NOME_ANTIGO_DO_OBJETO]
 gs://[NOVO_NOME_DO_OBJETO]
- D. gcloud mv gs://[NOME_ANTIGO_DO_OBJETO]
 gs://[NOVO_NOME_DO_OBJETO]

20. Um executivo da sua empresa lhe envia um e-mail perguntando sobre a criação de um sistema de recomendação que ajudará a vender mais produtos. O executivo ouviu dizer que existem algumas soluções do Google Cloud que podem ser adequadas para esse problema. Qual serviço do Google Cloud você recomendaria que o executivo investigasse?

- A. Cloud Dataproc, especialmente Spark e sua biblioteca de aprendizado de máquina
- B. Cloud Dataproc, especialmente Hadoop
- C. Cloud Spanner, que é um banco de dados relacional global capaz de armazenar uma grande quantidade de dados
- D. Cloud SQL, porque SQL é uma linguagem de consulta poderosa

Capítulo 13

Carregando Dados em Armazenamento

ESTE CAPÍTULO COBRE OS SEGUINTESS OBJETIVOS DO EXAME DE CERTIFICAÇÃO DE ENGENHEIRO ASSOCIADO DA GOOGLE CLOUD:

- ✓✓ 3.4 Implantando e implementando soluções de dados

Neste capítulo, mergulharemos nos detalhes de carregar e mover dados para vários sistemas de armazenamento e processamento no Google Cloud. Começaremos explicando como carregar e mover dados no Cloud Storage usando o console e a linha de comando. A maior parte do capítulo descreverá como importar e exportar dados para serviços de armazenamento e análise de dados, incluindo Cloud SQL, Cloud Firestore, BigQuery, Cloud Spanner, Cloud Bigtable e Cloud Dataproc. O capítulo termina com uma visão sobre o streaming de dados para o Cloud Pub/Sub.

Carregando e Movendo Dados para o Cloud Storage

O Cloud Storage é usado para uma variedade de casos de uso de armazenamento, incluindo armazenamento de longo prazo e arquivamento, transferências de arquivos e compartilhamento de dados. Esta seção descreve como criar buckets de armazenamento, carregar dados para buckets de armazenamento e mover objetos entre buckets de armazenamento.

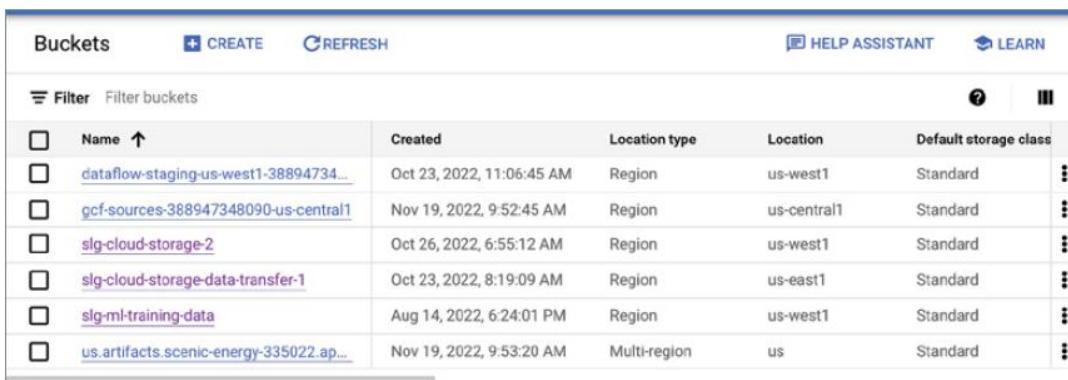
O Google Cloud recentemente introduziu o comando gcloud storage, que tem funcionalidades semelhantes ao gsutil. O gcloud storage geralmente é mais performático que o gsutil.

Carregando e Movendo Dados para o Cloud Storage Usando o Console

Carregar dados para o Cloud Storage é uma tarefa comum que é facilmente realizada usando o Cloud Console. Navegue até a página do Cloud Storage no Cloud Console. Você verá uma lista de buckets existentes e uma opção para criar um novo bucket. A Figura 13.1 mostra uma listagem de buckets e o botão Criar Bucket acima da lista.

Quando você cria um bucket, é solicitado que especifique um nome e um local onde deseja armazenar seus dados, como mostrado na Figura 13.2. O nome do bucket deve ser globalmente único. A localização pode ser Multi-Região para a maior disponibilidade e maior custo, Dual-Região para alta disponibilidade e baixa latência entre duas regiões, ou Região, que tem a menor latência dentro de uma única região. Se você escolher Multi-Região, suas opções incluem Estados Unidos, Europa e Ásia-Pacífico (consulte o console para a lista mais recente de multi-regiões). Se você escolher Dual-Região, pode especificar duas regiões dentro de um continente, com as opções atuais sendo Estados Unidos, Europa e Ásia-Pacífico. Se você escolher Região, então pode escolher qualquer uma das regiões disponíveis.

FIGURE 13.1 The first step in loading data into Cloud Storage is to create a bucket.



The screenshot shows the Google Cloud Storage Buckets page. At the top, there are buttons for 'CREATE' and 'REFRESH'. Below that is a 'HELP ASSISTANT' and 'LEARN' link. A 'Filter' button and a 'Filter buckets' input field are also present. The main area displays a table of existing buckets:

Name	Created	Location type	Location	Default storage class	Actions
dataflow-staging-us-west1-38894734...	Oct 23, 2022, 11:06:45 AM	Region	us-west1	Standard	⋮
gcf-sources-388947348090-us-central1	Nov 19, 2022, 9:52:45 AM	Region	us-central1	Standard	⋮
slg-cloud-storage-2	Oct 26, 2022, 6:55:12 AM	Region	us-west1	Standard	⋮
slg-cloud-storage-data-transfer-1	Oct 23, 2022, 8:19:09 AM	Region	us-east1	Standard	⋮
slg-ml-training-data	Aug 14, 2022, 6:24:01 PM	Region	us-west1	Standard	⋮
us.artifacts.scenic-energy-335022.ap...	Nov 19, 2022, 9:53:20 AM	Multi-region	us	Standard	⋮

FIGURE 13.2 Defining a regional bucket in us-west1

The screenshot shows the 'Name your bucket' step of a bucket creation wizard. It includes fields for naming the bucket, optional labels, and choosing a location type. The 'Region' option is selected for the location.

Name your bucket
Pick a globally unique, permanent name. [Naming guidelines](#)

ace-exam-test-bucket

Tip: Don't include any sensitive information

LABELS (OPTIONAL)

CONTINUE

Choose where to store your data

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. [Learn more](#)

Location type

Multi-region
Highest availability across largest area

Dual-region
High availability and low latency across 2 regions

Region
Lowest latency within a single region

us-west1 (Oregon)

CONTINUE

Lembre-se de que os buckets são recursos regionais e são replicados entre zonas na região.

A seguir, você precisará escolher sua classe de armazenamento (veja a Figura 13.3). As opções são Standard, que é melhor para armazenamento de curto prazo e objetos frequentemente acessados; Nearline para objetos acessados menos de uma vez a cada 30 dias; Coldline para objetos acessados menos de uma vez a cada 90 dias; e Arquivo, que é usado para objetos acessados menos de uma vez por ano.

FIGURE 13.3 Choosing a storage class and access control method

The screenshot shows a step in the Google Cloud Storage bucket creation process. The first section, "Choose a default storage class for your data," is active, indicated by a checked checkbox. It contains four options: "Standard" (selected), "Nearline," "Coldline," and "Archive." Each option has a brief description and a "Learn more" link. A "CONTINUE" button is at the bottom. The second section, "Choose how to control access to objects," is partially visible below it.

Choose a default storage class for your data

A storage class sets costs for storage, retrieval, and operations. Pick a default storage class based on how long you plan to store your data and how often it will be accessed. [Learn more](#)

Standard ?
Best for short-term storage and frequently accessed data

Nearline
Best for backups and data accessed less than once a month

Coldline
Best for disaster recovery and data accessed less than once a quarter

Archive
Best for long-term digital preservation of data accessed less than once a year

CONTINUE

Choose how to control access to objects

Prevent public access

Restrict data from being publicly accessible via the internet. Will prevent this bucket from being used for web hosting. [Learn more](#)

Enforce public access prevention on this bucket

Access control

Uniform
Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days. [Learn more](#)

Fine-grained
Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM). [Learn more](#)

CONTINUE

Aqui você também precisará decidir como deseja controlar o acesso ao bucket. Como os buckets são endereçáveis pela web, você pode permitir que qualquer pessoa com a URL do seu bucket acesse o conteúdo desse bucket. O Google Cloud oferece a opção de impedir explicitamente esse tipo de acesso público fornecendo a opção Importar Prevenção de Acesso Público Neste Bucket.

Originalmente, o Google Cloud usava listas de controle de acesso nos buckets para gerenciar o acesso a eles. Isso agora é chamado de opção de acesso detalhado, e permite especificar controles de acesso em objetos individuais, bem como em buckets. Embora o acesso detalhado ainda esteja disponível, a opção preferida é usar o acesso uniforme, no qual o acesso aos objetos no bucket é controlado por permissões no nível

do bucket gerenciadas pelo serviço IAM. O controle de acesso uniforme é o padrão e usar isso é considerado uma melhor prática.

Os exames de certificação do Google Cloud podem testar seu conhecimento sobre as práticas recomendadas pelo Google. Usar o controle de acesso uniforme em vez do controle de acesso detalhado é uma dessas práticas recomendadas.

Após criar um bucket, você pode visualizar os detalhes do bucket, conforme mostrado na Figura 13.4.

FIGURE 13.4 The Bucket Details page shows information on Objects, Configuration, Permissions, Protection, and Lifecycle.

The screenshot shows the 'Bucket details' page for the bucket 'ace-exam-test-bucket'. At the top, it displays basic bucket metadata: Location (us-east1 (South Carolina)), Storage class (Standard), Public access (Not public), and Protection (None). Below this, there are five tabs: OBJECTS (selected), CONFIGURATION, PERMISSIONS, PROTECTION, and LIFECYCLE. Under the OBJECTS tab, there are several buttons: UPLOAD FILES, UPLOAD FOLDER, CREATE FOLDER, MANAGE HOLDS, DOWNLOAD, and DELETE. There is also a 'Filter' button and a 'Show deleted data' checkbox. At the bottom, a message says 'No rows to display'.

Quando você clica em Carregar Arquivos, é solicitado a fazer isso usando o sistema de arquivos do seu dispositivo cliente. Quando você faz o upload de uma pasta, também é solicitado pelas ferramentas do sistema operacional local (veja a Figura 13.5).

É fácil mover objetos entre buckets. Basta clicar nos três pontos no final de uma linha para exibir uma lista de operações, que inclui Mover. Clicar em Mover abrirá a página mostrada na Figura 13.6.

Ao mover um objeto, é solicitado um bucket de destino e pasta, conforme mostrado na Figura 13.7.

FIGURE 13.5 Upload Files prompts you for a folder using the client device's filesystem tools.

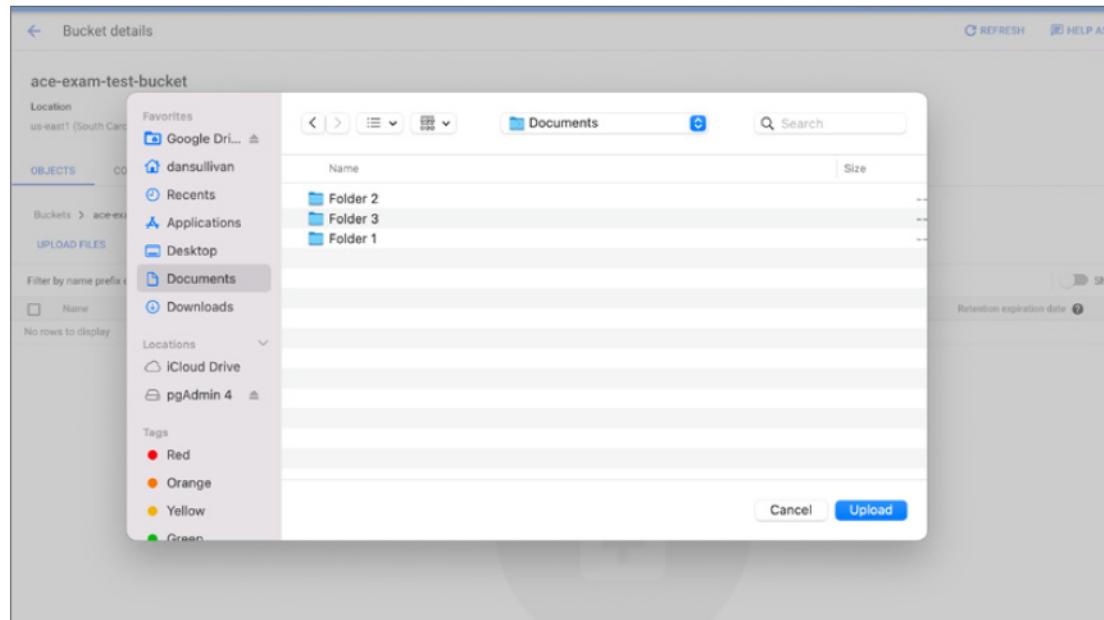


FIGURE 13.6 Objects can be moved by using the move command in the Operations menu.

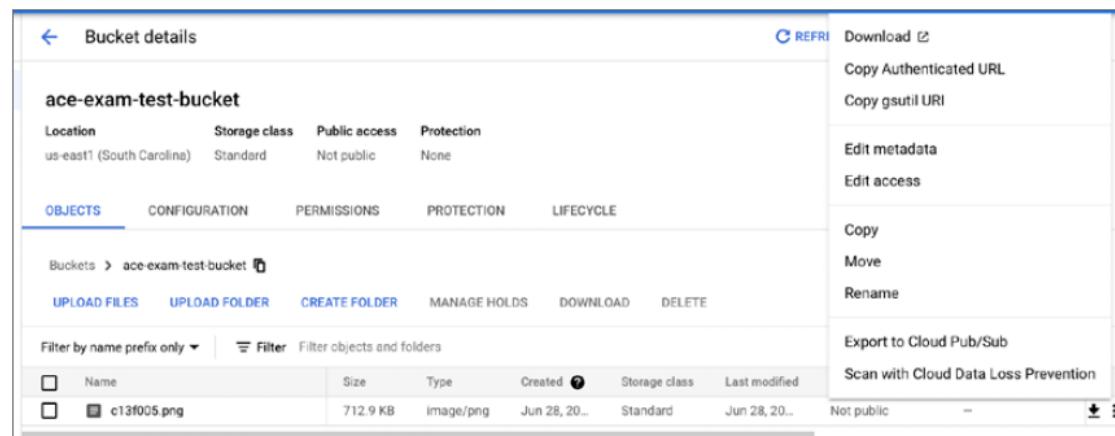
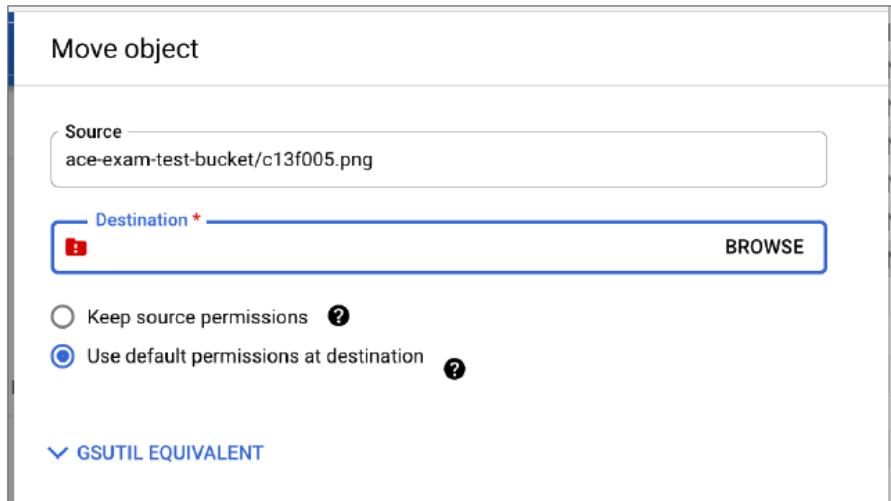


FIGURE 13.7 When moving an object in the console, you will be prompted for a destination bucket and folder.



Carregando e Movendo Dados para o Cloud Storage Usando a Linha de Comando

Carregar e mover dados pode ser feito na linha de comando usando o comando gsutil. Para criar um bucket, use o comando gsutil mb. "mb" é a abreviação de "make bucket" (criar bucket).

```
gsutil mb gs://[NOME_DO_BUCKET]/
```

Lembre-se de que os nomes dos buckets devem ser globalmente únicos. Para criar um bucket chamado ace-exam-bucket1, use o seguinte comando:

```
gsutil mb gs://ace-exam-bucket1/
```

Para fazer upload de um arquivo do seu dispositivo local ou de uma máquina virtual (VM) do Google Cloud, você pode usar o comando gsutil cp para copiar arquivos. O comando é o seguinte:

```
gsutil cp [LOCALIZAÇÃO_DO_OBJETO_LOCAL]  
gs://[NOME_DO_BUCKET_DE_DESTINO]/
```

Por exemplo, para copiar um arquivo chamado README.txt de /home/mydir para o bucket ace-exam-bucket1, você executaria o seguinte comando a partir da linha de comando do seu dispositivo cliente:

```
gsutil cp /home/mydir/README.txt gs://ace-exam-bucket1/
```

Similarmente, se você quiser baixar uma cópia dos seus dados de um bucket do Cloud Storage para um diretório em uma VM, você poderia fazer login na VM usando SSH e emitir um comando como este:

```
gsutil cp gs://ace-exam-bucket1/README.txt /home/mydir/
```

Neste exemplo, o objeto de origem está no Cloud Storage, e o arquivo de destino está na VM de onde você está executando o comando.

A ferramenta gsutil tem um comando de mover; sua estrutura é a seguinte:

```
gsutil mv  
gs://[NOME_DO_BUCKET_DE_ORIGEM]/[NOME_DO_OBJETO_DE_ORIGEM] \\\ngs://[NOME_DO_BUCKET_DE_DESTINO]/[NOME_DO_OBJETO_DE_DESTINO]
```

Para mover o arquivo README.txt de ace-exam-bucket1 para ace-exam-bucket2 e manter o mesmo nome de arquivo, você usaria este comando:

```
gsutil mv gs://ace-exam-bucket1/README.txt gs://ace-exam-bucket2/
```

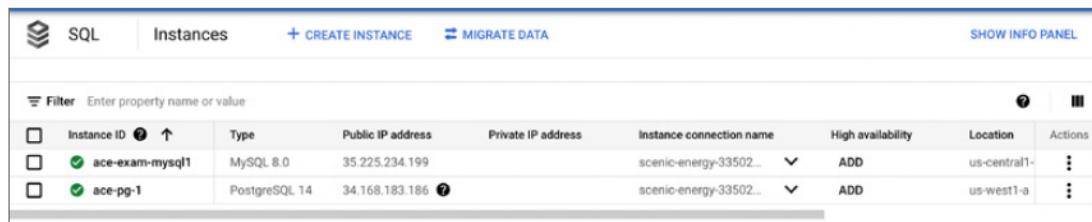
Importando e Exportando Dados

Como Engenheiro de Cloud, você pode precisar realizar operações em massa de dados, como importar e exportar dados de bancos de dados. Essas operações são feitas com ferramentas de linha de comando e, às vezes, o console. Não entraremos em detalhes sobre como inserir dados programaticamente em bancos de dados; isso é mais uma tarefa de desenvolvedor de aplicativos e administrador de banco de dados.

Importando e Exportando Dados: Cloud SQL

Para exportar um banco de dados Cloud SQL usando o console, navegue até a página do Cloud SQL do console para listar as instâncias de banco de dados, conforme mostrado na Figura 13.8.

FIGURE 13.8 Listing of database instances on the Cloud SQL page of the console



Instance ID	Type	Public IP address	Private IP address	Instance connection name	High availability	Location	Actions
<input type="checkbox"/> ace-exam-mysql1	MySQL 8.0	35.225.234.199		scenic-energy-33502...	▼ ADD	us-central1-a	⋮
<input type="checkbox"/> ace-pg-1	PostgreSQL 14	34.168.183.186		scenic-energy-33502...	▼ ADD	us-west1-a	⋮

Abra a página de Detalhes da Instância clicando duas vezes no nome da instância (veja a Figura 13.9). Selecione a aba Exportar para abrir a página Exportar Dados. Você precisará especificar um bucket no qual armazenar o arquivo de backup (veja a Figura 13.10). Você também precisará escolher entre a saída SQL ou CSV. A saída SQL é útil se você planeja importar os dados para outro banco de dados relacional. CSV é uma boa escolha se você precisa mover esses dados para um banco de dados não relacional ou outra ferramenta que não seja um banco de dados relacional. Após criar um arquivo de exportação, você pode importá-lo. Siga as mesmas instruções que para exportar, mas escolha a opção Importar em vez da opção Exportar. Isso exibirá as opções mostradas na Figura 13.11. Especifique o arquivo de origem, o formato do arquivo e o banco de dados para o qual importar os dados.

FIGURE 13.9 The Instance Details page has Import and Export tabs.

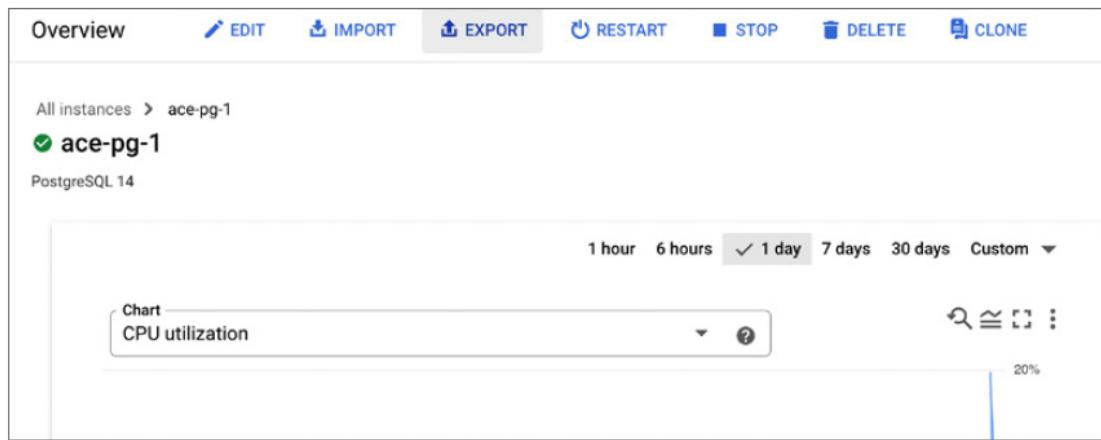


FIGURE 13.10 Exporting a database requires you to specify a bucket for storing the export file and a file format.

Export data to Cloud Storage

Source
Choose the format for your export, and the data you'd like to export from this instance.
[Learn more](#)

File format

SQL
A plain text file with a sequence of SQL commands, like the output of pg_dump.

CSV
Exports a plain text file with one line per row and comma-separated fields. Requires SQL SELECT query.

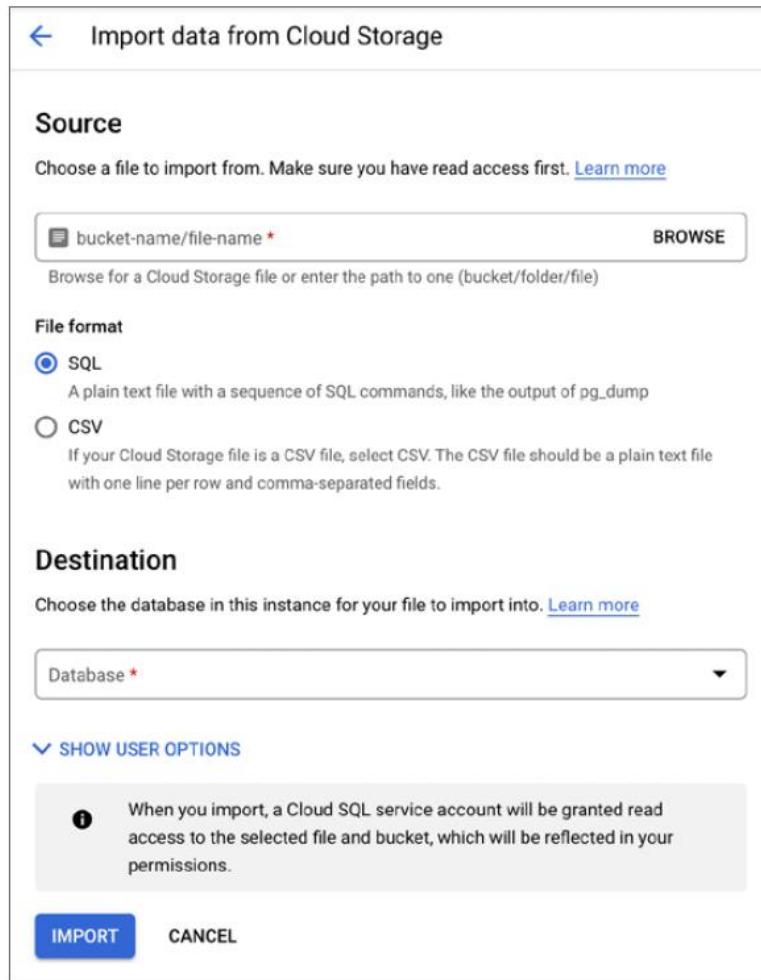
Data to export
Choose a database from this instance to export.

Offload export to a temporary instance
Makes your export serverless to reduce strain on the source instance, allowing you to perform other operations while the export is in progress. Affects cost. [Learn more](#)

Destination
Choose a Cloud Storage location to export into. Make sure you have the required permissions. [Learn more](#)

Browse for a Cloud Storage location or enter the path to one

FIGURE 13.11 Importing a database requires you to specify a path to the bucket and object storing the export file, a file format, and a target database within the instance.



Você também pode criar, importar e exportar um banco de dados usando a linha de comando. Use o comando gsutil para criar um bucket:

```
gsutil mb gs://ace-exam-bucket1/
```

Você precisa garantir que a conta de serviço possa escrever no bucket, então obtenha o nome da conta de serviço descrevendo a instância com o seguinte comando:

```
gcloud sql instances describe [NOME_DA_INSTÂNCIA]
```

Neste exemplo, o comando seria o seguinte:

```
gcloud sql instances describe ace-exam-mysql1
```

Este comando produzirá uma listagem detalhada sobre a instância. Veja a Figura 13.12 para um exemplo da saída.

FIGURE 13.12 Details about a database instance generated by the gcloud sql instances describe command

```

dan@cloudshell:~ (scenic-energy-335022)$ gcloud sql instances describe ace-exam-mysql1
backendType: SECOND_GEN
connectionName: scenic-energy-335022:us-central1:ace-exam-mysql1
createTime: '2022-06-30T00:32:53.562Z'
databaseInstalledVersion: MYSQL_8_0_26
databaseVersion: MYSQL_8_0
etag: 152a1dc634a450414ale93798b9d00074b6634d07ac4f4aa789d8b45e39ba251
gceZone: us-central1-f
instanceType: CLOUD_SQL_INSTANCE
ipAddresses:
- ipAddress: 35.225.234.199
  type: PRIMARY
kind: sql#instance
maintenanceVersion: MYSQL_8_0_26.R20220508.01_03
name: ace-exam-mysql1
project: scenic-energy-335022
region: us-central1
selfLink: https://sqladmin.googleapis.com/sql/v1beta4/projects/scenic-energy-335022/instances/ace-exam-mysql1
serverCaCert:
cert: |-
-----BEGIN CERTIFICATE-----
MIIDfzCCAmegAwIBAgIBADANBgkqhkiG9w0BAQsFADB3MS0wKwYDVQQQuEyRhZWNm
ZTVhNC0zYzowLTO4NiAtYTMwOC1hYzNlYzJlMzaxODYxTzAhBaNVBAMTGkdvh2ds
-----END CERTIFICATE-----

```

Você pode criar uma exportação de um banco de dados usando este comando:

```

gcloud      sql      export      sql      [NOME_DA_INSTÂNCIA]
gs://[NOME_DO_BUCKET]/[NOME_DO_ARQUIVO]          --
database=[NOME_DO_BANCO_DE_DADOS]

```

Por exemplo, o seguinte comando exportará o banco de dados MySQL para um arquivo de dump SQL escrito no bucket ace-exam-bucket1:

```

gcloud sql export sql ace-exam-mysql1 \
gs://ace-exam-bucket1/ace-exam-mysqlexport.sql --database=mysql

```

Se preferir exportar para um arquivo CSV, você mudará sql para csv no comando anterior. Aqui está um exemplo:

```

gcloud sql export csv ace-exam-mysql1 \
gs://ace-exam-bucket1/ace-exam-mysql-export.csv --database=mysql

```

A importação para um banco de dados usa um comando estruturado de maneira semelhante:

```

gcloud sql import sql [NOME_DA_INSTÂNCIA] \
gs://[NOME_DO_BUCKET]/[NOME_DO_ARQUIVO_DE_IMPORTAÇÃO]          --
database=[NOME_DO_BANCO_DE_DADOS]

```

Usando o banco de dados de exemplo, bucket e arquivo de exportação, você pode importar o arquivo usando este comando:

```

gcloud sql import sql ace-exam-mysql1 \
gs://ace-exam-bucket1/ace-exam-mysql-export.sql --database=mysql

```

Importando e Exportando Dados: Cloud Firestore

Para exportar dados do Cloud Firestore no modo Nativo, você pode usar este comando:

```
gcloud firestore export gs://${BUCKET}
```

A importação e exportação de dados do Firestore no modo Datastore é realizada através da linha de comando. O modo Datastore usa uma estrutura de dados de namespace para agrupar entidades que são exportadas. Você precisará especificar o nome do namespace usado pelas entidades que está exportando. O namespace padrão é simplesmente (default).

O comando de exportação do Cloud Datastore é o seguinte:

```
gcloud datastore export --namespaces="(default)" gs://${BUCKET}
```

Você pode exportar para um bucket chamado ace-exam-datastore1 usando este comando:

```
gcloud datastore export --namespaces="(default)" gs://ace-exam-datastore1
```

O comando de importação do Cloud Datastore é o seguinte:

```
gcloud datastore import gs://${BUCKET}/[Caminho]/[Arquivo].overall_export_metadata
```

O processo de exportação criará uma pasta chamada ace-exam-datastore1 usando a data e hora da exportação. A pasta conterá um arquivo de metadados e uma pasta contendo os dados exportados. O nome do arquivo de metadados usará a mesma data e hora usada para a pasta contendo. A pasta de dados será nomeada após o namespace do banco de dados Datastore exportado. Um exemplo de comando de importação é o seguinte:

```
gcloud datastore import gs://ace-exam-datastore1/2018-12-20T19:13:55_64324/2018-12-20T19:13:55_64324.overall_export_metadata
```

Importando e Exportando Dados: BigQuery

Usuários do BigQuery podem exportar e importar tabelas usando o Cloud Console e a linha de comando. Para exportar uma tabela usando o console, navegue até a interface do console do BigQuery. Em Recursos, abra o conjunto de dados que contém a tabela que você deseja exportar. Clique no nome da tabela para listar a descrição da tabela, conforme mostrado na Figura 13.13. Observe a opção Exportar no canto superior direito.

No extremo direito, clique em Exportar para exibir uma lista de quatro locais de exportação: Google Sheets, Google Cloud Storage, Looker Studio (anteriormente Data Studio), que é uma ferramenta de análise no Google Cloud, ou Scanning com o serviço de Prevenção de Perda de Dados (veja a Figura 13.14).

Selecionar Cloud Storage exibe as opções mostradas na Figura 13.15. Insira o nome do bucket no qual você deseja armazenar o arquivo de exportação. Escolha um formato de arquivo. As opções são CSV, Avro e JSON. Escolha um tipo de compressão. As opções são Nenhuma ou Gzip para CSV e Deflate e Snappy para Avro.

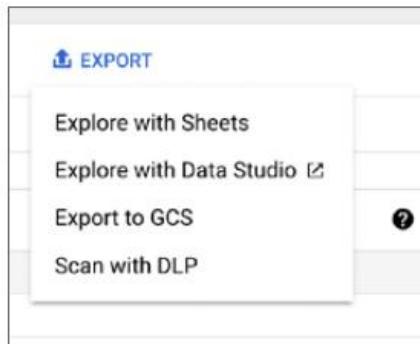
FIGURE 13.13 Detailed list of a BigQuery table

The screenshot shows the Google Cloud BigQuery interface. On the left, there's an 'Explorer' sidebar with pinned projects like 'scenic-energy-335022' and 'example_tables'. Under 'example_tables', 'mydata' is selected. The main area shows the 'SCHEMA' tab for the 'mydata' table. It contains three columns: 'Field name', 'Type', and 'Mode'. The rows are:

Field name	Type	Mode
id	INTEGER	NULLABLE
description	STRING	NULLABLE
status	STRING	NULLABLE

At the bottom, there are buttons for 'EDIT SCHEMA' and 'VIEW ROW ACCESS POLICIES'.

FIGURE 13.14 Choosing a target location for a BigQuery export



Formatos de Arquivo

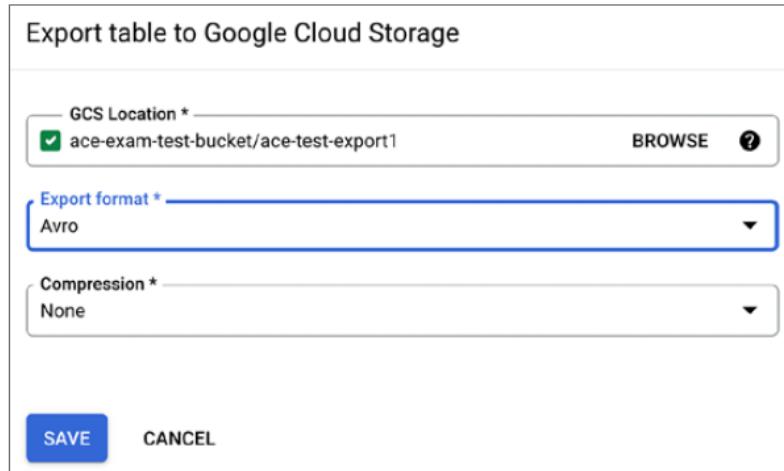
O BigQuery oferece várias opções de arquivo para exportação. CSV, abreviação para valores separados por vírgula, é um formato legível por humanos adequado para conjuntos de dados pequenos que serão importados para ferramentas que suportam apenas o formato CSV. O CSV não é otimizado para armazenamento, então ele não comprime ou usa uma codificação mais eficiente do que o texto. Não é a melhor opção ao exportar grandes conjuntos de dados.

JSON é também um formato legível por humanos que tem vantagens e desvantagens similares ao CSV. Uma diferença é que o JSON inclui informações de esquema com cada registro, enquanto o CSV usa uma linha de cabeçalho opcional com nomes de colunas no início do arquivo para descrever o esquema.

Gzip é uma utilidade de compressão sem perda amplamente utilizada.

Avro é um formato binário compacto que suporta estruturas de dados complexas. Quando os dados são salvos no formato Avro, um esquema é escrito no arquivo junto com os dados. Os esquemas são definidos em JSON. Avro é uma boa opção para grandes conjuntos de dados, especialmente ao importar dados para outras aplicações que leem o formato Avro, incluindo Apache Spark, que está disponível como um serviço gerenciado no Cloud Dataproc. Arquivos Avro podem ser comprimidos usando as utilidades deflate ou Snappy. Deflate produz arquivos comprimidos menores, mas Snappy é mais rápido.

FIGURE 13.15 Specifying the output parameters for a BigQuery export operation



Para exportar dados da linha de comando, use o comando `bq extract`. A estrutura é a seguinte:

```
bq extract --destination_format [FORMATO] --compression
[TIPO_DE_COMPRESSÃO] --field_delimiter [DELIMITADOR] --print_header
[BOOLEANO] [ID_DO_PROJETO]:[CONJUNTO_DE_DADOS].[TABELA]
gs://[BUCKET]/[NOME_DO_ARQUIVO]
```

Aqui está um exemplo:

```
bq extract --destination_format CSV --compression GZIP 'meudataset.minhatabela'
gs://exemplo-bucket/meuarquivo.zip
```

Lembre-se, a ferramenta de linha de comando para trabalhar com o BigQuery é `bq`, não `gcloud`.

Para importar dados para o BigQuery, navegue até a página do console do BigQuery e selecione um conjunto de dados para o qual você gostaria de importar dados. Clique em um conjunto de dados e, em seguida, selecione Criar Tabela, conforme mostrado na Figura 13.16.

A página Criar Tabela tem vários parâmetros, incluindo uma tabela de origem opcional, um projeto de destino, o nome do conjunto de dados, o tipo da tabela e o nome da tabela (veja a Figura 13.17).

O campo Criar Tabela De indica onde encontrar os dados de origem, se houver. Este campo fornece uma maneira de criar uma tabela baseada em dados em uma tabela existente, mas por padrão é uma tabela vazia (veja a Figura 13.18).

Você também precisará especificar o formato do arquivo que será importado. As opções incluem CSV, JSONL (JSON Delimitado por Nova Linha), Avro, Parquet, ORC e Backup do Cloud Datastore (veja a Figura 13.19).

FIGURE 13.16 When viewing a data set, you have the option to create a table.

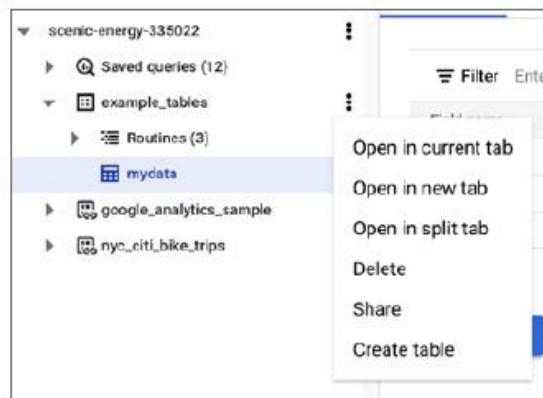


FIGURE 13.17 Creating a table in BigQuery

A screenshot of the 'Create table' dialog box. The dialog is divided into sections: 'Source' (with 'Create table from' dropdown showing 'Empty table'), 'Destination' (with 'Project' set to 'scenic-energy-335022', 'Dataset' set to 'example_tables', and 'Table' input field containing 'mydata2' which is underlined in red), and 'Schema' (with 'Edit as text' toggle button and a plus sign for adding fields). At the bottom are 'CREATE TABLE' and 'CANCEL' buttons.

FIGURE 13.18 Data can be imported from multiple kinds of locations.

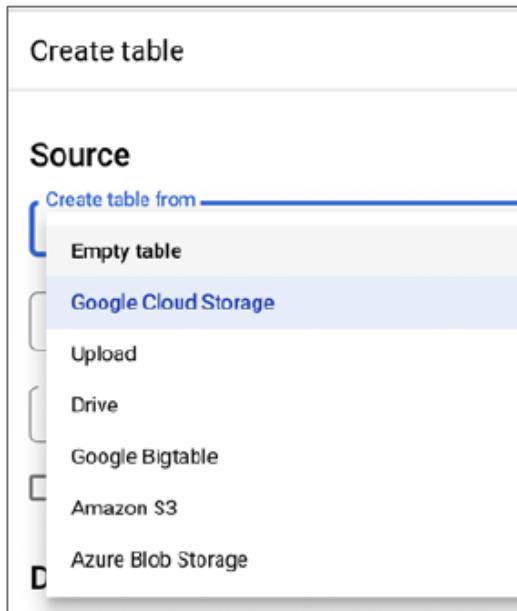
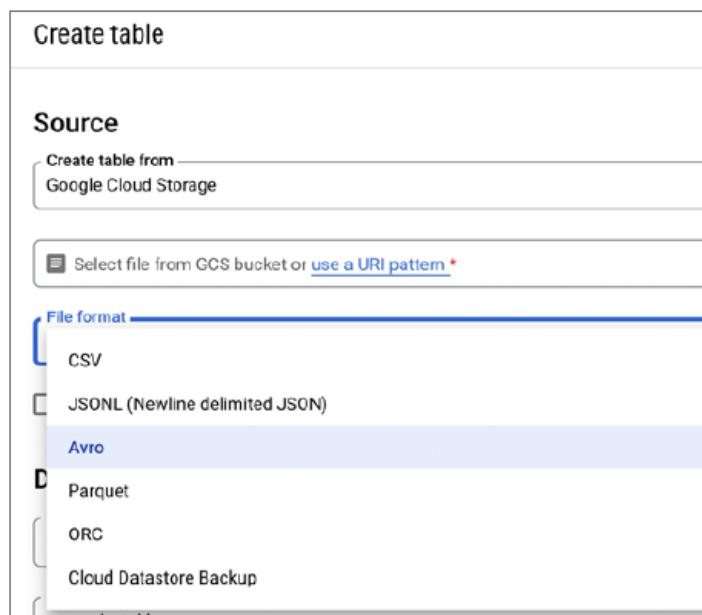


FIGURE 13.19 File format options for importing



Forneça informações de destino, incluindo projeto, nome do conjunto de dados, tipo de tabela e nome da tabela. O tipo de tabela pode ser do tipo Nativo ou uma tabela externa. Se a tabela for externa, os dados são mantidos no local de origem, e apenas metadados sobre a tabela são armazenados no BigQuery. Esse tipo é usado quando você tem grandes conjuntos de dados e não deseja carregar todos eles para o BigQuery.

Após especificar todos os parâmetros, clique em Criar Tabela para criar a tabela e carregar os dados.

Para carregar dados da linha de comando, use o comando bq load. Sua estrutura é a seguinte:

```
bq          load      --autodetect      --source_format=[FORMATO]
[CONJUNTO_DE_DADOS].[TABELA] [Caminho_PARA_FONTE]
```

O parâmetro --autodetect faz com que o bq load detecte automaticamente o esquema da tabela a partir do arquivo de origem. Um exemplo de comando é o seguinte:

```
bq load --autodetect --source_format=CSV meudataset.minhatabela gs://ace-exam-
biquery/meusdados.csv
```

Importando e Exportando Dados: Cloud Spanner

Usuários do Cloud Spanner podem importar e exportar dados usando o Cloud Console.

Para exportar dados do Cloud Spanner, navegue até a seção do Cloud Spanner no console. Você verá uma lista de instâncias do Spanner, conforme mostrado na Figura 13.20.

FIGURE 13.20 Listing of Spanner instances

Name	ID	Configuration	Processing units	Nodes	Storage utilization	Labels
ace-spanner-1	ace-spanner-1	us-west1 (Oregon)	100	0	0 B / 410 GB	

Clique no nome da instância que é a fonte dos dados a serem exportados. Isso mostrará a página de Detalhes da Instância (veja a Figura 13.21).

Clique em Exportar para exibir as opções de Exportação, conforme mostrado na Figura 13.22. Você precisará inserir um bucket de destino, o banco de dados a exportar e uma região para executar o trabalho. Observe que você deve confirmar que entende que haverá cobranças por executar o Cloud Dataflow e que pode haver cobranças de egresso de dados para dados enviados entre regiões.

Para importar dados, clique em Importar para exibir a página de Importação (veja a Figura 13.23). Você precisará especificar um bucket de origem, um banco de dados de destino e uma região para executar um trabalho.

FIGURE 13.21 Import/Export page

Spanner

All instances > INSTANCE ace-spanner-1: Import/Export

INSTANCE

Import/Export

In the past week, no import/export jobs have been run through the Cloud Spanner UI. To view all jobs for this project, [go to Cloud Dataflow](#).

Cloud Spanner's import/export feature assumes you intend to use an auto mode VPC network named `default` in the same project for the Dataflow job it creates. If you don't have a default VPC network in the project, or if your VPC network is custom mode, you need to directly create a Dataflow job and [specify a network and subnetwork](#).

IMPORT EXPORT

FIGURE 13.22 Export options for Cloud Spanner

← Export data from ace-spanner-1

Use this workflow to export data from a Cloud Spanner database into a Google Cloud Storage bucket. Your database will export in the form of a folder containing Apache Avro files. [Learn more](#)

Before you get started: Cloud Spanner imports use multiple Cloud Platform products. Make sure you have the required [permissions and/or quota](#) in Cloud Spanner, Cloud Storage, Compute Engine, and Cloud Dataflow.

1 **Choose where to store your export**

Destination ace-exam-test-bucket

2 **Choose a database to export**

Select a Cloud Spanner database to export into your Cloud Storage bucket.

Database name *

NEXT

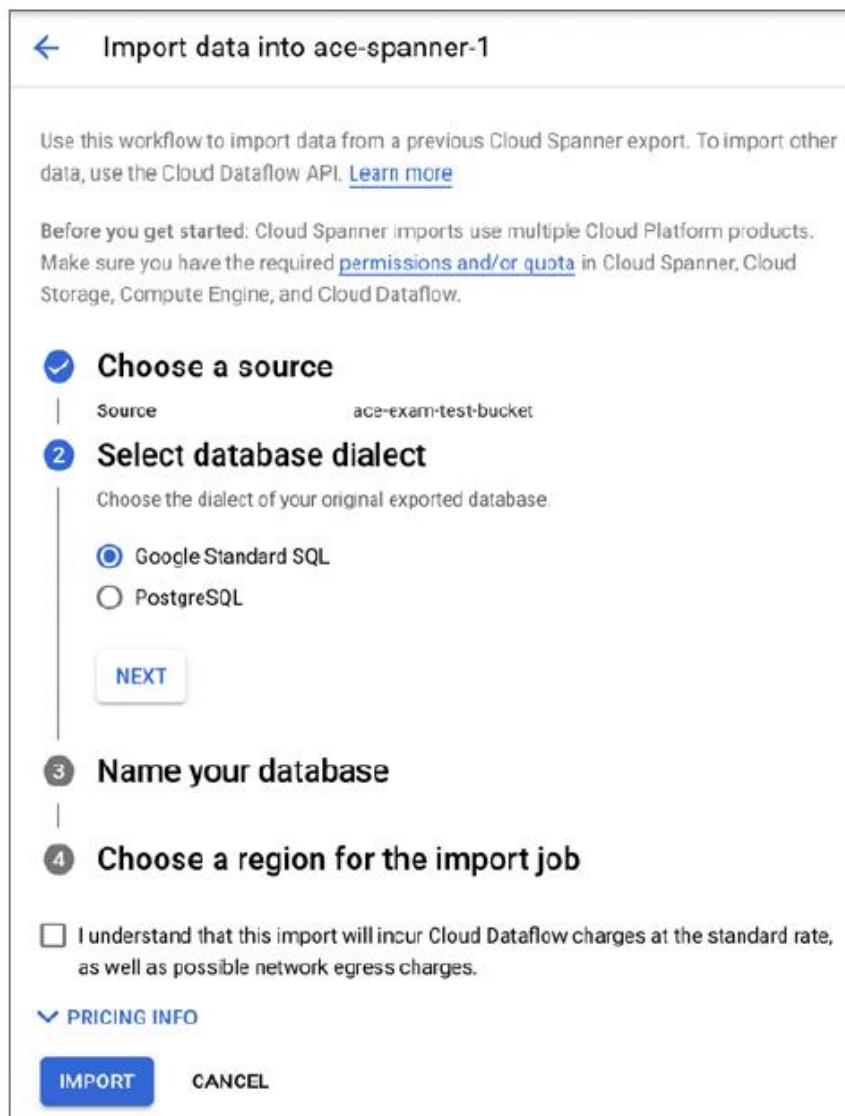
3 **Choose a region for the export job**

I understand that this export will incur Cloud Dataflow charges at the standard rate, as well as possible network egress charges.

PRICING INFO

EXPORT CANCEL

FIGURE 13.23 Import options for Cloud Spanner



O Cloud Spanner não possui um comando gcloud para exportar dados, mas você pode usar o Dataflow para exportar dados. Os detalhes de construção de trabalhos do Dataflow estão fora do escopo desta seção. Para mais detalhes, consulte a documentação do Cloud Dataflow em <https://cloud.google.com/dataflow/docs>.

Exportando Dados do Cloud Bigtable

O Cloud Bigtable suporta a exportação de dados através do console. Navegue até o console do Bigtable e selecione Tabelas na barra de menu à esquerda. Isso exibirá um diálogo de exportação conforme mostrado na Figura 13.24. As exportações do Bigtable são armazenadas no Cloud Storage e podem usar um dos três formatos: SequenceFile, Avro ou Parquet.

FIGURE 13.24 Export page for Cloud Bigtable

Table ID	Cluster	Status	Storage utilization	Latest timestamp
ace-bigtable-table-1	ace-bigtable-1	✓	SequenceFile Files on Cloud Storage Avro Files on Cloud Storage Parquet Files on Cloud Storage	

Importando e Exportando Dados: Cloud Dataproc

O Cloud Dataproc não é um banco de dados como o Cloud SQL ou Bigtable; é uma plataforma de análise de dados. Essas plataformas são projetadas mais para manipulação de dados, análise estatística, aprendizado de máquina e outras operações complexas do que para armazenamento e recuperação de dados. O Cloud Dataproc não é projetado para ser um armazenamento persistente de dados. Para isso, você deve usar o Cloud Storage ou discos persistentes para armazenar os arquivos de dados que deseja analisar.

O Cloud Dataproc possui comandos de Importar e Exportar para salvar e restaurar dados de configuração de cluster. Esses comandos estão disponíveis usando o gcloud.

O comando para exportar a configuração de um cluster Dataproc é o seguinte:

```
gcloud dataproc clusters export [NOME_DO_CLUSTER] --destination=[CAMINHO PARA ARQUIVO DE EXPORTAÇÃO]
```

Aqui está um exemplo:

```
gcloud dataproc clusters export ace-exam-dataproc-cluster --destination=gs://ace-exam-bucket1/meudataprocyaml
```

Para importar um arquivo de configuração, use o comando de importação:

```
gcloud dataproc clusters import [ARQUIVO_DE_ORIGEM]
```

Por exemplo, para importar o arquivo criado no exemplo de exportação anterior, use o seguinte:

```
gcloud dataproc clusters import gs://ace-exam-bucket1/meudataprocyaml
```

Importar e exportar dados são operações comuns. O Google Cloud oferece ferramentas de console e de linha de comando para a maioria dos serviços de banco de dados. Existem também comandos em beta para exportar e importar dados de configuração de cluster para o Dataproc.

Transmitindo Dados para o Cloud Pub/Sub

Até agora neste capítulo, você passou a maior parte do seu tempo movendo dados para dentro e ao redor do Cloud Storage, além de importar e exportar dados para bancos de dados. Agora, vamos voltar nossa atenção para trabalhar com o Cloud Pub/Sub, a fila de mensagens.

Como Engenheiro de Cloud, você pode precisar criar filas de mensagens para desenvolvedores de aplicativos. Embora seja provável que os desenvolvedores escrevam serviços que usem o Pub/Sub, os Engenheiros de Cloud devem ser capazes de testar tópicos e inscrições do Pub/Sub. Discutimos como criar filas de mensagens no Capítulo 12, “Implantando Armazenamento no Google Cloud.” Aqui, nosso foco será na criação de mensagens em tópicos e recebendo essas mensagens por meio de inscrições.

Os comandos do gcloud pubsub que você usará são create, publish e pull. Para criar um tópico, use o seguinte comando:

```
gcloud pubsub topics create [NOME_DO_TÓPICO]
```

O comando para criar uma inscrição é o seguinte:

```
gcloud pubsub subscriptions create --topic [NOME_DO_TÓPICO] [NOME_DA_INSCRIÇÃO]
```

Por exemplo, para criar um tópico chamado ace-exam-topic1 e uma inscrição para esse tópico chamada ace-exam-sub1, você pode usar estes comandos:

```
gcloud pubsub topics create ace-exam-topic1
```

```
gcloud pubsub subscriptions create --topic=ace-exam-topic1 ace-exam-sub1
```

Agora, para testar se a fila de mensagens está funcionando corretamente, você pode enviar dados para o tópico usando o seguinte comando:

```
gcloud pubsub topics publish [NOME_DO_TÓPICO] --message [MENSAGEM]
```

e depois ler essa mensagem da inscrição usando o seguinte:

```
gcloud pubsub subscriptions pull --auto-ack [NOME_DA_INSCRIÇÃO]
```

Para escrever uma mensagem no tópico e lê-la da inscrição que você acabou de criar, você pode usar o seguinte:

```
gcloud pubsub topics publish ace-exam-topic1 --message "primeira mensagem do exame ace"
```

```
gcloud pubsub subscriptions pull --auto-ack ace-exam-sub1
```

Mundo Real

Desacoplando Serviços Usando Filas de Mensagens

Um dos desafios com sistemas distribuídos é que às vezes um serviço não consegue acompanhar o fluxo de dados. Isso pode criar um backlog em serviços que dependem do serviço atrasado.

Por exemplo, um pico repentino de tráfego em um site de varejo pode colocar uma alta carga em um serviço de rastreamento de inventário, que atualiza o inventário conforme os clientes adicionam ou removem itens de seus cestos. O programa de inventário pode ser lento para responder a um serviço que adicionou um item ao carrinho. Se esse serviço estiver esperando uma resposta do serviço de inventário, ele também será atrasado. Esse tipo de comunicação síncrona é problemático quando sistemas distribuídos estão sob carga.

Uma opção melhor é desacoplar a conexão direta entre os serviços. Por exemplo, a interface do usuário poderia escrever uma mensagem em um tópico Pub/Sub toda vez que um item é adicionado ou removido do cesto de um cliente. O serviço de gerenciamento de inventário pode se inscrever neste tópico e atualizar o sistema de inventário à medida que novas mensagens chegam. Se o sistema de inventário desacelerar, não afetará a interface do usuário porque ela está escrevendo em um tópico Pub/Sub, que pode escalar junto com a carga gerada pela interface do usuário.

Resumo

Neste capítulo, examinamos as diferentes maneiras de carregar dados em armazenamento, bancos de dados e sistemas de fila de mensagens. O Cloud Storage é organizado em torno de objetos em buckets. O comando gsutil e o Cloud Console podem ser usados para fazer upload de dados, bem como movê-los entre buckets. Você viu que o comando gsutil cp pode ser usado para copiar arquivos entre o Cloud Storage e VMs.

Os serviços de banco de dados fornecem utilitários de importação e exportação. Cada um suporta uma variedade de formatos de arquivo.

O Cloud Pub/Sub pode ser usado para desacoplar aplicações e melhorar a resiliência a picos de carga. Você viu como criar um tópico e inscrições e como empurrar dados para a fila de mensagens, onde pode ser lido por inscritos.

Saiba que o Cloud Spanner usa o serviço Dataflow para importar e exportar. Pode haver cobranças adicionais ao usar o Dataflow e mover dados entre regiões.

Essenciais para o Exame

Saiba como carregar dados e movimentar dados no Cloud Storage. O Cloud Storage é amplamente utilizado para uma variedade de casos de uso, incluindo armazenamento de longo prazo e arquivamento, transferências de arquivos e compartilhamento de dados. Entenda a estrutura dos comandos gsutil, que é diferente de gcloud. Os comandos gsutil começam com gsutil seguido por uma operação, como copiar ou criar bucket. Certifique-se de conhecer a sintaxe dos comandos de cópia (cp), mover (mv) e criar bucket (mb). Você pode copiar arquivos do Cloud Storage para VMs e vice-versa. Além disso, saiba que o comando gsutil acl ch -u é usado para alterar permissões em objetos. Você pode usar o comando gsutil acl ch para alterar permissões em um bucket do Cloud Storage.

Entenda como funcionam a importação e exportação com o Cloud SQL. Importar e exportar dados de bancos de dados são operações comuns. Você pode realizar importações e exportações a partir do console e da linha de comando.

Saiba que você pode exportar entidades de um Cloud Firestore. Exportações e importações são feitas no nível do banco de dados quando no modo Nativo e no nível de namespaces quando o banco de dados está no modo Datastore.

Entenda como exportar e importar dados do BigQuery. O BigQuery tem uma gama de opções para a fonte de dados a importar. Os dados podem ser comprimidos quando exportados para economizar espaço. O BigQuery pode exportar dados em vários formatos, incluindo CSV, JSON e Avro. Saiba que o comando `bq` é usado para importar e exportar da linha de comando.

Saiba que o Pub/Sub é usado para enviar mensagens entre serviços. O Pub/Sub permite maior resiliência às flutuações de carga. Se um serviço atrasa, seu trabalho pode acumular em uma fila Pub/Sub sem forçar o serviço que gera esses dados a esperar.

Questões de Revisão

Você pode encontrar as respostas no Apêndice.

1. Qual dos seguintes comandos é usado para criar buckets no Cloud Storage?
 - A. gcloud storage create buckets
 - B. gsutil storage buckets create
 - C. gsutil mb
 - D. gcloud mb
2. Você precisa copiar arquivos do seu dispositivo local para um bucket no Cloud Storage. Que comando você usaria? Assuma que você tenha o Cloud SDK instalado no seu computador local.
 - A. gsutil copy
 - B. gsutil cp
 - C. gcloud cp
 - D. gcloud storage objects copy
3. Você está migrando um grande número de arquivos de um sistema de armazenamento local para o Cloud Storage. Você quer usar o Cloud Console ao invés de escrever um script. Qual das seguintes operações do Cloud Storage você pode realizar no console?
 - A. Apenas fazer upload de arquivos
 - B. Apenas fazer upload de pastas
 - C. Fazer upload de arquivos e pastas
 - D. Comparar arquivos locais com arquivos no bucket usando o comando diff
4. Um desenvolvedor de software pede sua ajuda para exportar dados de um banco de dados Cloud SQL. O desenvolvedor lhe diz qual banco de dados exportar e em qual bucket armazenar o arquivo de exportação, mas não mencionou qual formato de arquivo deve ser usado para o arquivo de exportação. Quais são as opções para o formato do arquivo de exportação?
 - A. CSV e XML
 - B. CSV e JSON
 - C. JSON e SQL
 - D. CSV e SQL
5. Um administrador de banco de dados pediu por uma exportação de um banco de dados MySQL no Cloud SQL. O administrador de banco de dados irá carregar os

dados em outro banco de dados relacional e gostaria de fazer isso com o mínimo de trabalho possível. Especificamente, o método de carga não deve requerer que o administrador de banco de dados defina um esquema. Qual formato de arquivo você recomendaria para esta tarefa?

- A. SQL
 - B. CSV
 - C. XML
 - D. JSON
6. Qual comando exportará um banco de dados MySQL chamado ace-exam-mysql11 para um arquivo chamado ace-exam-mysql-export.sql em um bucket chamado ace-exam-bucketel?
- A. gcloud storage export sql ace-exam-mysql11
gs://ace-exam-bucketel/ace-exam-mysql-export.sql
—database=mysql
 - B. gcloud sql export ace-exam-mysql11
gs://ace-exam-bucketel/ace-exam-mysql-export.sql
—database=mysql
 - C. gcloud sql export sql ace-exam-mysql11
gs://ace-exam-bucketel/ace-exam-mysql-export.sql
—database=mysql
 - D. gcloud sql export sql ace-exam-mysql11
gs://ace-exam-mysql-export.sql/ace-exam-bucketel/
—database=mysql
7. Qual dos seguintes formatos de arquivo não é uma opção para um arquivo de exportação ao exportar do BigQuery?
- A. CSV
 - B. XML
 - C. Avro
 - D. JSON
8. Qual dos seguintes formatos de arquivo não é suportado ao importar dados para o BigQuery?
- A. CSV
 - B. Parquet

- C. Avro
 - D. YAML
9. Você recebeu um grande conjunto de dados de um sistema de Internet das Coisas (IoT). Você quer usar o BigQuery para analisar os dados. Que comando de linha de comando você usaria para tornar os dados disponíveis para análise no BigQuery?
- A. bq load —autodetect —source_format=[FORMATO] [CONJUNTO_DE_DADOS].[TABELA] [Caminho_PARA_FONTE]
 - B. bq import —autodetect —source_format=[FORMATO] [CONJUNTO_DE_DADOS].[TABELA] [Caminho_PARA_FONTE]
 - C. gcloud BigQuery load —autodetect —source_format=[FORMATO] [CONJUNTO_DE_DADOS].[TABELA] [Caminho_PARA_FONTE]
 - D. gcloud BigQuery load —autodetect —source_format=[FORMATO] [CONJUNTO_DE_DADOS].[TABELA] [Caminho_PARA_FONTE]
10. Você configurou um processo no Cloud Spanner para exportar dados para o Cloud Storage. Você nota que, cada vez que o processo é executado, você incorre em cobranças de outro serviço do Google Cloud, o qual você acha que está relacionado ao processo de exportação. Que outro serviço do Google Cloud pode estar incorrendo em cobranças durante a exportação do Cloud Spanner?
- A. Dataproc
 - B. Dataflow
 - C. Firestore
 - D. bq
11. Como desenvolvedor em um projeto usando Bigtable para uma aplicação IoT, você precisará exportar dados do Bigtable para disponibilizar alguns dados para análise com outra ferramenta. O que você usaria para exportar os dados, assumindo que você deseja minimizar a quantidade de esforço necessário de sua parte?
- A. Um programa Java projetado para importar e exportar dados do Bigtable
 - B. gcloud bigtable table export
 - C. bq bigtable table export
 - D. Uma ferramenta de importação fornecida pela ferramenta de análise
12. Você acabou de exportar de um cluster Dataproc. O que você exportou?
- A. Dados em Spark DataFrames

- B. Todas as tabelas no banco de dados Spark
 - C. Dados de configuração sobre o cluster
 - D. Todas as tabelas no banco de dados Hadoop
13. Uma equipe de cientistas de dados solicitou acesso a dados armazenados no Bigtable para que possam treinar modelos de aprendizado de máquina. Eles explicam que o Bigtable não possui os recursos necessários para construir modelos de aprendizado de máquina. Qual dos seguintes serviços do Google Cloud eles provavelmente usarão para construir modelos de aprendizado de máquina?
- A. Firestore
 - B. Dataflow
 - C. Dataproc
 - D. DataAnalyze
14. Qual dos seguintes é o comando correto para criar um tópico no Pub/Sub?
- A. gcloud pubsub topics create
 - B. gcloud pubsub create topics
 - C. bq pubsub create topics
 - D. cbt pubsub topics create
15. Qual dos seguintes comandos criará uma inscrição no tópico ace-exam-topic1?
- A. gcloud pubsub create —topic=ace-exam-topic1 ace-exam-sub1
 - B. gcloud pubsub subscriptions create —topic=ace-exam-topic1
 - C. gcloud pubsub subscriptions create —topic=ace-exam-topic1 ace-exam-sub1
 - D. gsutil pubsub subscriptions create —topic=ace-exam-topic1 ace-exam-sub1
16. Qual é uma das vantagens diretas de usar uma fila de mensagens em sistemas distribuídos?
- A. Aumenta a segurança.
 - B. Desacopla serviços, então se um atrasar, não faz com que outros serviços atrasem.
 - C. Suporta mais linguagens de programação.
 - D. Armazena mensagens até que sejam lidas por padrão.
17. Para garantir que você instalou comandos beta do gcloud, qual comando você deve executar?
- A. gcloud components beta install

- B. gcloud components install beta
 - C. gcloud commands install beta
 - D. gcloud commands beta install
18. Qual parâmetro é usado para dizer ao BigQuery para detectar automaticamente o esquema de um arquivo na importação?
- A. autodetect
 - B. autoschema
 - C. detectschema
 - D. dry_run
19. As opções de compressão Deflate e Snappy estão disponíveis para quais tipos de arquivo ao exportar do BigQuery?
- A. Avro
 - B. CSV
 - C. XML
 - D. Thrift
20. Você quer ler uma mensagem de um tópico Pub/Sub e reconhecer a leitura dessa mensagem no mesmo comando. Qual dos seguintes você usaria?
- A. gcloud pubsub subscriptions pull --auto-ack
 - B. gcloud pubsub topic pull --auto-ack
 - C. gsutil pubsub topic pull --with-acknowledgement
 - D. gcloud pubsub subscription pull --with-acknowledgement

Capítulo 14

Redes na Nuvem: Nuvens Privadas Virtuais e Redes Privadas Virtuais

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO DE ENGENHEIRO ASSOCIADO DA GOOGLE CLOUD:

✓✓ 2.4 Planejamento e configuração de recursos de rede

✓✓ 4.5 Gerenciamento de recursos de rede

Neste capítulo, voltamos nossa atenção para a rede, começando com nuvens privadas virtuais (VPCs). Você aprenderá como criar VPCs com sub-redes padrão e personalizadas. Você aprenderá sobre a criação de configurações de rede personalizadas no Compute Engine para casos em que as configurações de rede padrão não atendem às suas necessidades. Finalmente, mostraremos como configurar regras de firewall e criar redes privadas virtuais (VPNs).

Criando uma Nuvem Privada Virtual com Sub-redes

VPCs são versões de software de redes físicas que ligam recursos em um projeto. O Google Cloud cria automaticamente uma VPC quando você cria um projeto. Você pode criar VPCs adicionais e modificar as VPCs criadas pelo Google Cloud.

VPCs são recursos globais, então não estão vinculados a uma região ou zona específica. Recursos, como máquinas virtuais (VMs) do Compute Engine e clusters do Kubernetes Engine, podem se comunicar entre si, assumindo que o tráfego não seja bloqueado por uma regra de firewall.

VPCs contêm sub-redes, chamadas subnets, que são recursos regionais. Subnets têm uma gama de endereços IP associados a elas. Subnets fornecem endereços internos privados. Recursos usam esses endereços para se comunicar entre si e com APIs e serviços do Google.

Além das VPCs associadas a projetos, você pode criar uma VPC compartilhada dentro de uma organização. A VPC compartilhada é hospedada em um projeto comum. Usuários em outros projetos que têm permissões suficientes podem criar e usar recursos na VPC compartilhada. Você também pode usar o emparelhamento de redes VPC para conectividade interprojetos, mesmo se uma organização não estiver definida.

Nesta seção, você criará uma VPC com subnets usando o Cloud Console e gcloud, e depois voltará sua atenção para a criação de uma VPC compartilhada.

Criando uma Nuvem Privada Virtual com o Cloud Console

Para criar uma VPC no Cloud Console, navegue até a página de Redes VPC, como mostrado na Figura 14.1.

Clicar em Criar Rede VPC abre a página mostrada na Figura 14.2. A Figura 14.2 mostra que você pode atribuir um nome e descrição a uma nova VPC. Ela também mostra uma lista de subnets que serão criadas na VPC. Quando uma VPC no modo automático é criada, subnets são criadas em cada região. O Google Cloud escolhe uma gama de endereços IP para cada subnet ao criar uma rede no modo automático.

FIGURE 14.1 The VPC Network page of Cloud Console

VPC networks		CREATE VPC NETWORK	REFRESH	HELP ASSISTANT			
NETWORKS IN CURRENT PROJECT		SUBNETS IN CURRENT PROJECT					
SMTP port 25 disallowed in this project							
VPC networks							
Filter Enter property name or value							
Name	Subnets	MTU	Mode	Internal IP ranges	Gateways		
default	36	1460	Auto		Firewall rules 4 Global dynamic routing Off		

FIGURE 14.2 Creating a VPC in Cloud Console, part 1

Create a VPC network

Name *	
Lowercase letters, numbers, hyphens allowed	
Description	

Subnets

Subnets let you create your own private cloud topology within Google Cloud. Click Automatic to create a subnet in each region, or click Custom to manually define the subnets. [Learn more](#)

Subnet creation mode

Custom Automatic

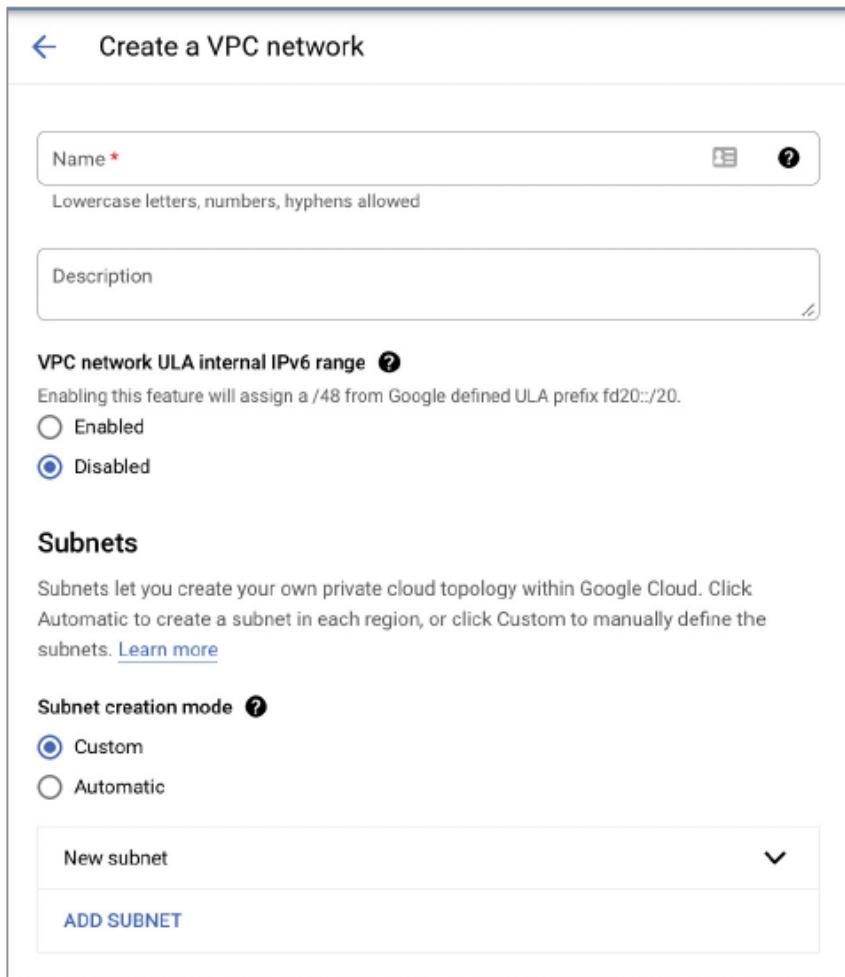
IP stack type
IPv4 (single-stack)

These IP address ranges will be assigned to each region in your VPC network. When an instance is created for your VPC network, it will be assigned an IP from the appropriate region's address range.

Region	IP address range
asia-east1	10.140.0.0/20
asia-northeast1	10.146.0.0/20
asia-south1	10.160.0.0/20
asia-southeast1	10.148.0.0/20
australia-southeast1	10.152.0.0/20

Alternativamente, você pode criar uma ou mais sub-redes personalizadas selecionando a aba Personalizado na seção de Subnet (Figura 14.3). Isso exibirá outra página que permite especificar uma região e uma faixa de endereços IP. A faixa de IP é especificada na notação de roteamento interdomínios sem classe (CIDR). (Veja o sidebar “Entendendo a Notação CIDR” para detalhes sobre como especificar endereços IP usando essa notação.) Você pode ativar o Acesso Privado Google. Isso permite que VMs na subnet accessem serviços Google sem atribuir um endereço IP externo à VM. Você também pode ativar o registro de tráfego de rede definindo a opção de Logs de Fluxo para Ativo.

FIGURE 14.3 Creating a custom subnet



A Figura 14.4 mostra a segunda parte da página da VPC, que inclui regras de firewall, configuração de roteamento dinâmico e uma política de servidor DNS. A seção Regras de Firewall lista regras que podem ser aplicadas à VPC. No exemplo da Figura 14.4, uma das regras permite entrada, que é tráfego TCP de entrada na porta 22, para permitir acesso SSH. A faixa de IP de 0.0.0.0/0 permite tráfego de todos os endereços IP de origem.

FIGURE 14.4 Creating a VPC in Cloud Console, part 2

The screenshot shows the 'Create a VPC network' interface. At the top, there's a back arrow and the title 'Create a VPC network'. Below that is a section for 'Firewall rules' with a note: 'Select any of the firewall rules below that you would like to apply to this VPC network. Once the VPC network is created, you can manage all firewall rules on the Firewall rules page.' There are two tabs: 'IPV4 FIREWALL RULES' (selected) and 'IPV6 FIREWALL RULES'. The table lists the following rules:

Name	Type	Targets	Filters	Protocols / ports	Action	Priority ↑	EDIT
allow-custom	Ingress	Apply to all	IP ranges:	all	Allow	65,534	EDIT
allow-icmp	Ingress	Apply to all	IP ranges:	icmp 0.0.0.0/0	Allow	65,534	
allow-rdp	Ingress	Apply to all	IP ranges:	tcp:3389 0.0.0.0/0	Allow	65,534	
allow-ssh	Ingress	Apply to all	IP ranges:	tcp:22 0.0.0.0/0	Allow	65,534	
deny-all-ingress	Ingress	Apply to all	IP ranges:	0.0.0.0/0	Deny	65,535	
allow-all-egress	Egress	Apply to all	IP ranges:	all 0.0.0.0/0	Allow	65,535	

Below the table is a 'Dynamic routing mode' section with two options: 'Regional' (selected) and 'Global'. The 'Regional' option is described as 'Cloud Routers will learn routes only in the region in which they were created'. The 'Global' option is described as 'Global routing lets you dynamically learn routes to and from all regions with a single VPN or interconnect and Cloud Router'. There is also a note: 'Enable DNS API to pick a DNS policy' with an 'ENABLE' button.

At the bottom, there is a dropdown for 'Maximum transmission unit (MTU)' set to '1460', and two buttons: 'CREATE' (highlighted in blue) and 'CANCEL'.

A opção de roteamento dinâmico determina quais rotas são aprendidas. O roteamento regional fará com que os Roteadores do Google Cloud aprendam rotas dentro da região. O roteamento global permitirá que os Roteadores do Google Cloud aprendam rotas em todas as sub-redes na VPC.

A política opcional do servidor DNS permite escolher uma política de DNS que possibilita a resolução de nomes DNS fornecida pelo Google Cloud ou faz mudanças na ordem de resolução de nomes. (Veja o Capítulo 15, “Redes na Nuvem: DNS, Balanceamento de Carga, Acesso Privado Google e Endereçamento IP,” para mais detalhes.)

Uma vez que você tenha especificado os parâmetros e criado uma VPC, ela aparecerá na listagem de VPCs e mostrará informações sobre a VPC e suas sub-redes, conforme mostrado na Figura 14.5.

FIGURE 14.5 Listing of VPCs and subnets

The screenshot shows a table listing VPCs and their subnets. The columns are: Name (sorted by name), Region, Stack Type, Internal IP ranges, External IP ranges, and Secondary IPv4 ranges. There are seven rows, each representing a VPC with its subnets. The first row is 'ace-vpc-1' in 'us-west2'. The second row is 'default' in 'us-central1'. The third row is 'default' in 'europe-west1'. The fourth row is 'default' in 'us-west1'. The fifth row is 'default' in 'asia-east1'. The sixth row is 'default' in 'us-east1'. The seventh row is another 'default' entry.

	Name	Region	Stack Type	Internal IP ranges	External IP ranges	Secondary IPv4 ranges
<input type="checkbox"/>	ace-vpc-1	us-west2	IPv4	10.10.0.0/24	None	None
<input type="checkbox"/>	default	us-central1	IPv4	10.128.0.0/20	None	None
<input type="checkbox"/>	default	europe-west1	IPv4	10.132.0.0/20	None	None
<input type="checkbox"/>	default	us-west1	IPv4	10.138.0.0/20	None	None
<input type="checkbox"/>	default	asia-east1	IPv4	10.140.0.0/20	None	None
<input type="checkbox"/>	default	us-east1	IPv4	10.142.0.0/20	None	None

Criando uma Nuvem Privada Virtual com gcloud

O comando gcloud para criar uma VPC é gcloud compute networks create. Por exemplo, para criar uma VPC no projeto padrão com sub-redes geradas automaticamente, você usaria o seguinte comando:

```
gcloud compute networks create ace-exam-vpc1 --subnet-mode=auto
```

Você também pode configurar sub-redes personalizadas criando uma rede VPC especificando a opção customizada e então criando sub-redes nessa VPC. O primeiro comando para criar uma VPC personalizada chamada ace-exam-vpc1 é o seguinte:

```
gcloud compute networks create ace-exam-vpc1 --subnet-mode=custom
```

Em seguida, você pode criar uma sub-rede usando o comando gcloud compute networks subnets create. Este comando exige que você especifique uma VPC, a região e a faixa de IP. Você pode opcionalmente ativar as configurações de Acesso Privado Google e Logs de Fluxo adicionando as bandeiras apropriadas.

Aqui está um exemplo de comando para criar uma sub-rede chamada ace-exam-vpc-subnet1 na VPC ace-exam-vpc1. Esta sub-rede é criada na região us-west2 com uma faixa de IP de 10.10.0.0/16. As configurações de Acesso IP Privado e Logs de Fluxo estão ativadas.

```
gcloud compute networks subnets create ace-exam-vpc-subnet1 --network=ace-exam-vpc1 --region=us-west2 --range=10.10.0.0/16 --enable-private-ip-google-access --enable-flow-logs
```

Entendendo a Notação CIDR

Quando você especifica faixas de endereços IP, utiliza algo chamado roteamento interdomínios sem classe (CIDR). O nome vem das primeiras redes IP que foram definidas em três classes fixas primárias: A, B e C. Uma estrutura de endereçamento de rede sem classe foi criada para superar as limitações de uma estrutura de roteamento baseada em classe, particularmente a falta de flexibilidade na criação de sub-redes de diferentes tamanhos.

O CIDR usa máscara de sub-rede de comprimento variável (VLSM) para permitir que os administradores de rede definam redes com o número de endereços de que precisam, e não os números fixos que foram alocados à rotina interdomínios do modelo de classe antigo.

Endereços CIDR consistem em dois conjuntos de números: um endereço de rede para identificar uma sub-rede e um identificador de host. Esses números são escritos usando a notação CIDR, que consiste em um endereço de rede e uma máscara de rede. Exemplos de endereços de rede, de acordo com a especificação RFC1918, são:

- 10.0.0.0 - 10.255.255.255 (/8)
- 172.16.0.0 - 172.31.255.255 (/12)
- 192.168.0.0 - 192.168.255.255 (/16)

A notação CIDR adiciona uma barra (/) e um número indicando quantos bits de um endereço IP são alocados para a máscara de rede, o que determina quais endereços estão dentro do bloco do endereço e quais não estão.

Por exemplo, 192.168.0.0/16 significa que 16 bits dos 32 bits de um endereço IP são usados para especificar a rede, e 16 bits são usados para especificar o endereço do host. Com 16 bits, você pode criar $2^{16}-2$, ou 65.534 endereços de host.

O bloco CIDR 172.16.0.0/12 indica que 12 bits são usados para especificar a rede, e 20 bits são usados para especificar endereços de host. Com 20 bits, você pode criar até 1.048.574 endereços de host. Em geral, quanto menor o número após a barra, mais endereços de host estão disponíveis. Você pode experimentar com opções de bloco CIDR usando uma calculadora CIDR, como a disponível em www.subnet-calculator.com/cidr.php.

Criando uma Nuvem Privada Virtual Compartilhada Usando gcloud

Se você deseja criar uma VPC compartilhada, pode usar o comando gcloud compute shared-vpc. Antes de executar comandos para criar uma VPC compartilhada, você precisará atribuir a um membro da organização a função de Administrador de VPC Compartilhada no nível da organização ou do diretório. Para atribuir a função de Administrador de VPC Compartilhada, que usa o descritor roles/compute.xpnAdmin, emita este comando:

```
gcloud organizations add-iam-policy-binding [ORG_ID] --  
member='user:[ENDEREÇO_DE_EMAIL]' --role="roles/compute.xpnAdmin"
```

[ORG_ID] é o identificador da organização que utiliza a política. Você pode encontrar um ID de organização com o comando gcloud organizations list. Se preferir atribuir a função de Administrador de VPC Compartilhada a um diretório, você pode usar este comando:

```
gcloud resource-manager folders add-iam-policy-binding [FOLDER_ID] --  
member='user:[ENDEREÇO_DE_EMAIL]' --role="roles/compute.xpnAdmin"
```

[FOLDER_ID] é o identificador do diretório da política. Você pode obter IDs de diretórios usando este comando:

```
gcloud resource-manager folders list --organization=[ORG_ID]
```

Para mais informações sobre funções e privilégios, veja o Capítulo 17, “Configurando Acesso e Segurança.”

Uma vez que você tenha definido a função de Administrador de VPC Compartilhada no nível da organização, você pode emitir o comando shared-vpc:

```
gcloud compute shared-vpc enable [HOST_PROJECT_ID]
```

Se você estiver compartilhando a VPC no nível de pasta, use este comando:

```
gcloud compute shared-vpc enable [HOST_PROJECT_ID]
```

Agora que a VPC compartilhada foi criada, você pode associar projetos usando o comando gcloud compute shared-vpc associate-projects. No nível da organização, você pode usar este comando:

```
gcloud compute shared-vpc associated-projects add [SERVICE_PROJECT_ID] --host-project [HOST_PROJECT_ID]
```

No nível de pasta, o comando para associar pastas é o seguinte:

```
gcloud compute shared-vpc associated-projects add [SERVICE_PROJECT_ID] --host-project [HOST_PROJECT_ID]
```

Alternativamente, o emparelhamento de rede VPC pode ser usado para tráfego interprojetos quando uma organização não existe. O emparelhamento de rede VPC é implementado usando o comando gcloud compute networks peerings create. Por exemplo, você emparelha duas VPCs especificando emparelhamentos em cada rede. Aqui está um exemplo:

```
gcloud compute networks peerings create peer-ace-exam-1 --network ace-exam-network-A --peer-project ace-exam-project-B --peer-network ace-exam-network-B --auto-create-routes
```

E então crie um emparelhamento na outra rede usando:

```
gcloud compute networks peerings create peer-ace-exam-1 --network ace-exam-network-B --peer-project ace-exam-project-A --peer-network ace-exam-network-A --auto-create-routes
```

Implantando o Compute Engine com uma Rede Personalizada

Você pode implantar uma VM com configurações de rede personalizadas usando o console e a linha de comando.

Navegue até a seção do Compute Engine no console e abra a página Criar Instância, mostrada na Figura 14.6.

FIGURE 14.6 Preliminary options to create an instance in Cloud Console

The screenshot shows the initial configuration steps for creating a new Google Compute Engine instance. It includes fields for:

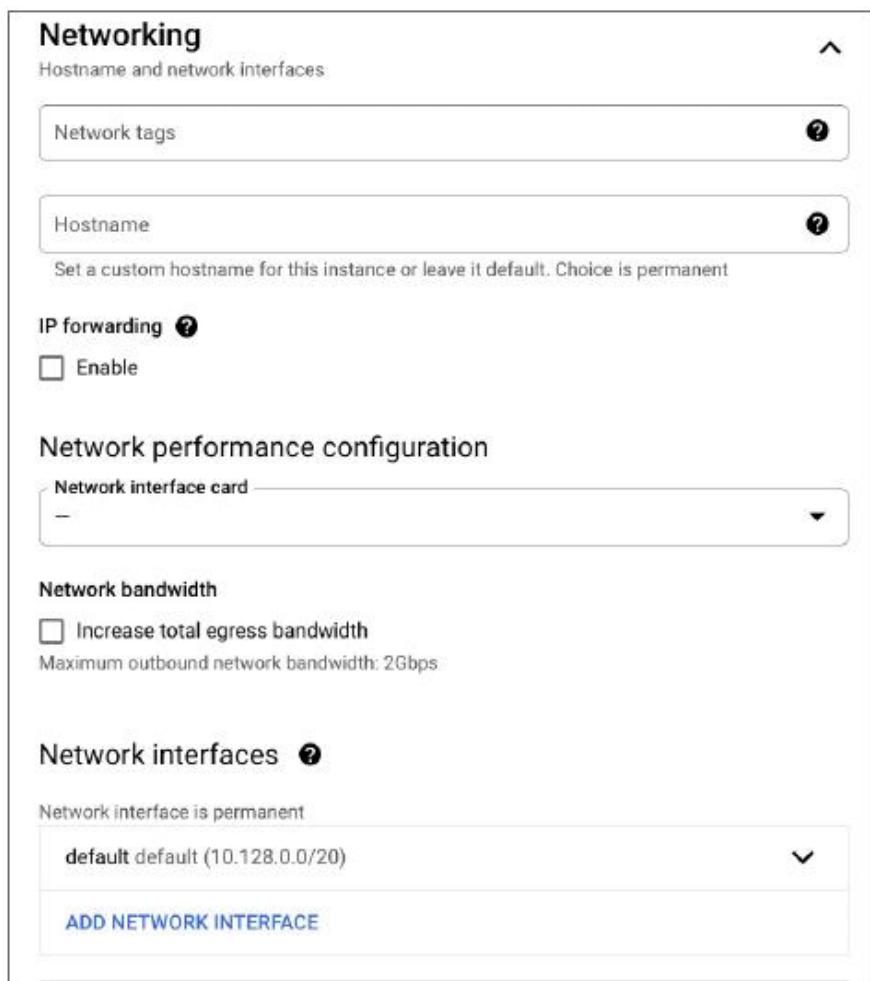
- Name ***: instance-1
- Labels**: A button to add labels.
- Region ***: us-central1 (Iowa)
- Zone ***: us-central1-a
- Machine configuration** section:
 - Machine family**: GENERAL-PURPOSE (selected)
 - Series**: E2
 - Machine type**: e2-medium (2 vCPU, 4 GB memory)
 - Hardware details**:

	vCPU	Memory
	1-2 vCPU (1 shared core)	4 GB
 - CPU PLATFORM AND GPU** (button)

No menu horizontal na parte inferior da página, clique em Gerenciamento > Segurança > Discos > Rede > Locação Exclusiva para expandir os formulários opcionais e, em seguida, clique na aba Rede para exibir uma página semelhante à Figura 14.7.

Observe que, nesta página, você pode definir tags de rede, que são usadas para definir regras de firewall e rotas. Clique em Adicionar Interface de Rede para exibir uma página como a mostrada na Figura 14.8. Aqui você pode escolher uma rede personalizada. Neste exemplo, estamos escolhendo ace-exam-vpc1, que criamos anteriormente no capítulo. Também selecionamos uma subnet.

FIGURE 14.7 Networking configuration options

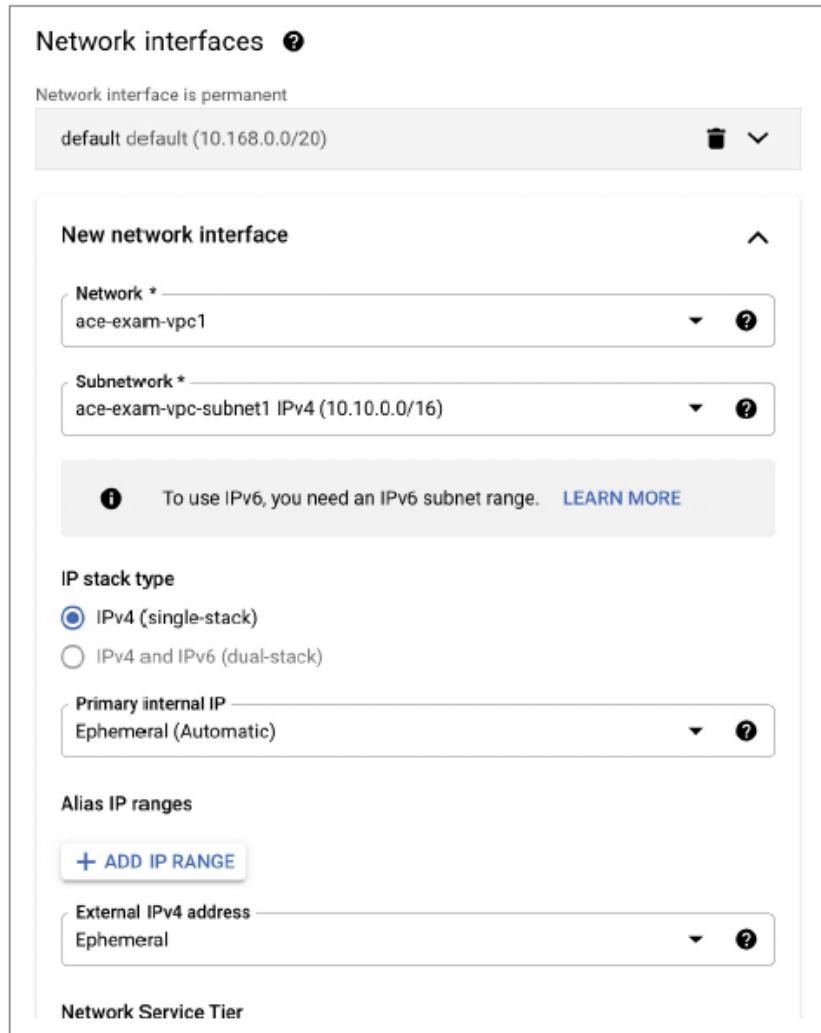


Aqui, você também pode especificar um endereço IP estático ou escolher um endereço efêmero personalizado usando a configuração de IP Interno Primário. O menu suspenso de IP Externo permite que você tenha um IP externo efêmero ou use um IP externo estático.

Você também pode criar uma instância para executar em uma subnet específica usando o comando gcloud compute instances create com parâmetros de subnet e zona.

```
gcloud compute instances create [NOME_DA_INSTÂNCIA] --subnet [NOME_DA_SUBNET] --zone [NOME_DA_ZONA]
```

FIGURE 14.8 Options to add a custom network interface



Criando Regras de Firewall para uma Nuvem Privada Virtual

As regras de firewall são definidas no nível da rede e usadas para controlar o fluxo de tráfego de rede para VMs.

As regras de firewall permitem ou negam um tipo especificado de tráfego em uma porta; por exemplo, uma regra pode permitir tráfego TCP para a porta 22. Elas também são aplicadas ao tráfego em uma direção, seja entrada (ingresso) ou saída (egresso). É importante notar que o firewall é stateful, o que significa que se o tráfego é permitido em uma direção e uma conexão estabelecida, ele é permitido na outra direção. Conjuntos de regras de firewall são stateful, então se uma conexão é permitida, como estabelecer uma conexão SSH na porta 22, então todo tráfego posterior que corresponda a esta regra é permitido enquanto a conexão estiver ativa. Uma conexão ativa é aquela com pelo menos um pacote trocado a cada 10 minutos.

Estrutura das Regras de Firewall

As regras de firewall consistem em vários componentes:

- Direção: Entrada (ingress) ou saída (egress).
- Prioridade: As regras de maior prioridade são aplicadas; qualquer regra com uma prioridade menor que corresponda não é aplicada. A prioridade é especificada por um inteiro de 0 a 65535. 0 é a maior prioridade e 65535 é a menor.
- Ação: Permitir (allow) ou negar (deny). Apenas uma pode ser escolhida.
- Alvo: Uma instância à qual a regra se aplica. Os alvos podem ser todas as instâncias em uma rede, instâncias com tags de rede particulares ou instâncias usando uma conta de serviço específica.
- Origem/Destino: Origem aplica-se a regras de entrada e especifica faixas de IP de origem, instâncias com tags de rede particulares ou instâncias usando uma conta de serviço específica. Você também pode usar combinações de faixas de IP de origem e tags de rede e combinações de faixas de IP de origem e contas de serviço usadas por instâncias. O endereço IP 0.0.0.0/0 indica qualquer endereço IP. O parâmetro Destino usa apenas faixas de IP.
- Protocolo e Porta: Um protocolo de rede como TCP, UDP ou ICMP e um número de porta. Se nenhum protocolo é especificado, então a regra se aplica a todos os protocolos.
- Status de Execução: As regras de firewall estão habilitadas ou desabilitadas. Regras desabilitadas não são aplicadas mesmo se corresponderem. Desabilitar é às vezes usado para solucionar problemas com tráfego que não passa quando deveria ou passa quando não deveria.
- Todas as VPCs começam com duas regras implícitas: uma permite tráfego de saída para todos os destinos (endereço IP 0.0.0.0/0) e uma nega todo o tráfego de entrada de qualquer fonte (endereço IP 0.0.0.0/0). Ambas as regras implícitas têm prioridade 65535, então você pode criar outras regras com prioridade maior para negar ou permitir tráfego conforme necessário. Você não pode excluir uma regra implícita.

Quando uma VPC é criada automaticamente, a rede padrão é criada com quatro regras de rede. Estas regras permitem o seguinte:

- Tráfego de entrada de qualquer instância de VM na mesma rede
- Tráfego TCP de entrada na porta 22, permitindo SSH
- Tráfego TCP de entrada na porta 3389, permitindo o Protocolo de Área de Trabalho Remota da Microsoft (RDP)
- Protocolo de Mensagem de Controle da Internet (ICMP) de entrada de qualquer fonte na rede

As regras padrão têm todas a prioridade 65534.

Criando Regras de Firewall Usando o Cloud Console

Para criar ou editar regras de firewall, navegue até a seção VPC do console e selecione a opção Firewall no menu VPC. A Figura 14.9 mostra uma lista de regras de firewall.

FIGURE 14.9 List of firewall rules in the VPC section of Cloud Console

	Name	Type	Targets	Filters	Protocols / ports	Action	Priority	Network	Logs	Hit count	Last hit	Insights
<input type="checkbox"/>	default-allow-icmp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	icmp	Allow	65534	default	Off	—	—	
<input type="checkbox"/>	default-allow-internal	Ingress	Apply to all	IP ranges: 10.0.0.0/16	tcp:0-65535 udp:0-65535 icmp	Allow	65534	default	Off	—	—	
<input type="checkbox"/>	default-allow-tcp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:3389	Allow	65534	default	Off	—	—	
<input type="checkbox"/>	default-allow-ssh	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:22	Allow	65534	default	Off	—	—	

Clique em Criar Regra de Firewall no topo da página para criar uma nova regra de firewall. Isso abre a página mostrada na Figura 14.10. Aqui, você especifica um nome e descrição da regra de firewall. Você pode escolher ativar ou desativar o registro. Se estiver ativado, as informações de registro serão capturadas no Cloud Logging. (Veja o Capítulo 18, “Monitoramento, Registro e Estimativa de Custos,” para mais sobre o Cloud Logging.) Você também precisa especificar a rede na VPC para aplicar a regra.

Em seguida, você precisará especificar uma prioridade, direção, ação, alvos e fontes. A prioridade pode ser números inteiros na faixa de 0 a 65535. A direção pode ser Entrada (Ingress) ou Saída (Egress). A ação pode ser Permitir (Allow) ou Negar (Deny). Escolha alvos na lista suspensa; as opções são mostradas na Figura 14.11.

Se você escolher tags ou contas de serviço, você poderá especificar as tags ou o nome da conta de serviço. Você também pode especificar filtros de fonte como faixas de IP, sub-redes, tags de fonte ou contas de serviço. O Google Cloud permite um segundo filtro de fonte se você quiser usar uma combinação de condições. Uma lista de filtros de fonte é mostrada na Figura 14.12.

Finalmente, você especifica protocolo e portas escolhendo entre as opções Permitir Tudo (Allow All) e Protocolos e Portas Especificados (Specified Protocols and Ports). Se você escolher a última, pode especificar protocolos e portas. A Figura 14.13 mostra a listagem da regra de firewall criada usando os parâmetros especificados na Figura 14.10.

FIGURE 14.10 Creating a firewall rule

[←](#) Create a firewall rule

Description
Example firewall rule

Logs
Turning on firewall logs can generate a large number of logs which can increase costs in Cloud Logging. [Learn more](#)

On
 Off

Network *
default

Priority *
1000 [CHECK PRIORITY OF OTHER FIREWALL RULES](#) [?](#)

Priority can be 0 - 65535

Direction of traffic [?](#)

Ingress
 Egress

Action on match [?](#)

Allow
 Deny

Targets
All Instances in the network

Source filter
IPv4 ranges

Source IPv4 ranges *
0.0.0.0/0 [\(x\)](#) for example, 0.0.0.0/0, 192.168.2.0/24 [?](#)

Second source filter
None

Protocols and ports [?](#)

Allow all
 Specified protocols and ports

[▼ DISABLE RULE](#)

CREATE **CANCEL**

FIGURE 14.11 List of target types

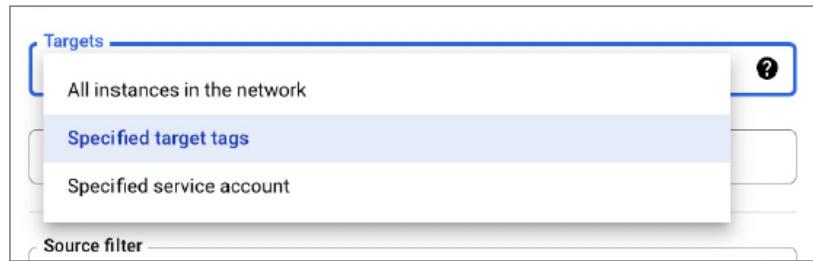


FIGURE 14.12 List of source filter types

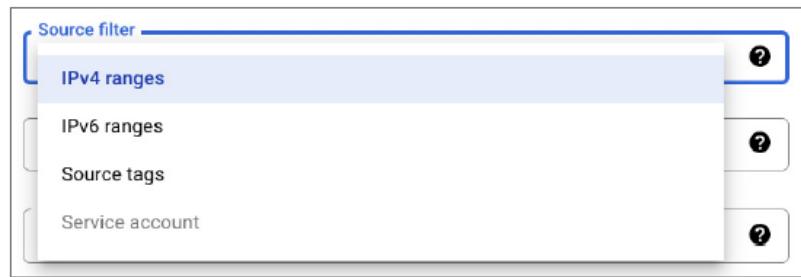


FIGURE 14.13 Listing of the firewall rule created using the earlier configuration

The screenshot shows the details of a firewall rule named 'ace-exam-fwr1'. The top bar includes a back arrow, the rule name, 'EDIT' and 'DELETE' buttons, and a help icon. The rule details are listed below:

- Logs**: Off, [view in Logs Explorer](#)
- Network**: default
- Priority**: 1000
- Direction**: Ingress
- Action on match**: Allow
- Targets**:
 - Target tags: ace-exam
- Source filters**:
 - IP ranges: 0.0.0.0/0

Criando Regras de Firewall Usando gcloud

O comando para trabalhar com regras de firewall a partir da linha de comando é `gcloud compute firewall-rules`. Com este comando, você pode criar, excluir, descrever, atualizar e listar regras de firewall.

Vários parâmetros são usados com `gcloud compute firewall-rules create`:

- `--action`
- `--allow`
- `--description`
- `--destination-ranges`
- `--direction`
- `--network`
- `--priority`
- `--source-ranges`
- `--source-service-accounts`
- `--source-tags`
- `--target-service-accounts`
- `--target-tags`

Por exemplo, para permitir todo o tráfego TCP nas portas de 20000 a 25000, use isto:

```
gcloud compute firewall-rules create ace-exam-fwr2 --network ace-exam-vpc1 --allow tcp:20000-25000
```

Criando uma Rede Privada Virtual

VPNs permitem enviar tráfego de rede de forma segura da rede do Google para sua própria rede. Você pode criar uma VPN usando o Cloud Console ou a linha de comando.

Criando uma Rede Privada Virtual Usando o Cloud Console

Para criar uma VPN usando o Cloud Console, navegue até a seção de Conectividade Híbrida do console, como mostrado na Figura 14.14.

Clique em Criar Conexão VPN para exibir a página mostrada na Figura 14.15.

Você tem a opção de criar uma VPN de Alta Disponibilidade (HA) ou uma VPN Clássica. VPNs HA suportam roteamento dinâmico usando o Protocolo de Gateway de Borda (BGP) bem como um SLA de alta disponibilidade de 99,99% dentro de uma região. A alta disponibilidade é fornecida usando dois túneis em vez de apenas um. Você pode usar endereços IPv4 ou IPv6 em uma VPN HA.

FIGURE 14.14 Hybrid Connectivity section of Cloud Console

The screenshot shows the 'Hybrid Connectivity' section in the Google Cloud console. On the left, there's a sidebar with four options: 'VPN' (selected), 'Interconnect', 'Cloud Routers', and 'Network Connectivity Center'. The main content area is titled 'VPN' and contains a brief description: 'A virtual private network lets you securely connect your Google Compute Engine resources to your own private network. Google VPN uses IKEv1 or IKEv2 to establish the IPSec connectivity.' Below the description is a blue 'CREATE VPN CONNECTION' button.

FIGURE 14.15 Creating a VPN connection, part 1

The screenshot shows the 'Create a VPN' page. At the top, there's a back arrow and the title 'Create a VPN'. Below the title is a description: 'A virtual private network lets you securely connect your Google Compute Engine resources to your own private network. Google VPN uses IKEv1 or IKEv2 to establish the IPSec connectivity.' Below the description are two options:

- High-availability (HA) VPN** (selected): Supports dynamic routing (BGP) only, Supports high availability (99.99 SLA, within region), Supports IPv4 and IPv6 traffic. [Learn more](#)
- Classic VPN**: Supports dynamic routing and static routing, No high availability, Supports IPv4 traffic only. [Learn more](#)

Below the options is a diagram illustrating the HA VPN setup. It shows an 'On-premise network' connected to a 'VPC network' via two tunnels: 'Tunnel1' and 'Tunnel2'. The 'VPC network' is represented by a light blue box containing a central router with two 'Gateway interface' ports, labeled 'Gateway interface0' and 'Gateway interface1'.

At the bottom of the page are 'CONTINUE' and 'CANCEL' buttons.

No passado, o Google Cloud oferecia uma VPN Clássica que suportava tanto o roteamento dinâmico quanto o estático, mas apenas endereços IPv4, e não fornecia alta

disponibilidade. A VPN Clássica foi parcialmente depreciada. Você pode continuar a usar túneis de VPN Clássica que usam roteamento dinâmico ao conectar-se a uma VM do Compute Engine executando um gateway VPN. Você não pode usar túneis de VPN Clássica para conexões fora do Google Cloud. (Veja <https://cloud.google.com/network-connectivity/docs/vpn/deprecations/classic-vpn-deprecation> para detalhes adicionais.)

A Figura 14.16 mostra a primeira parte da página para criar uma VPN HA. Você especifica um nome de gateway VPN, uma rede e uma região.

FIGURE 14.16 Creating a high availability VPN

The screenshot shows the 'Create a VPN' wizard, step 1: Create Cloud HA VPN gateway. The page has a header 'Create a VPN' with a back arrow. Below it, a section titled '1 Create Cloud HA VPN gateway' contains the following fields:

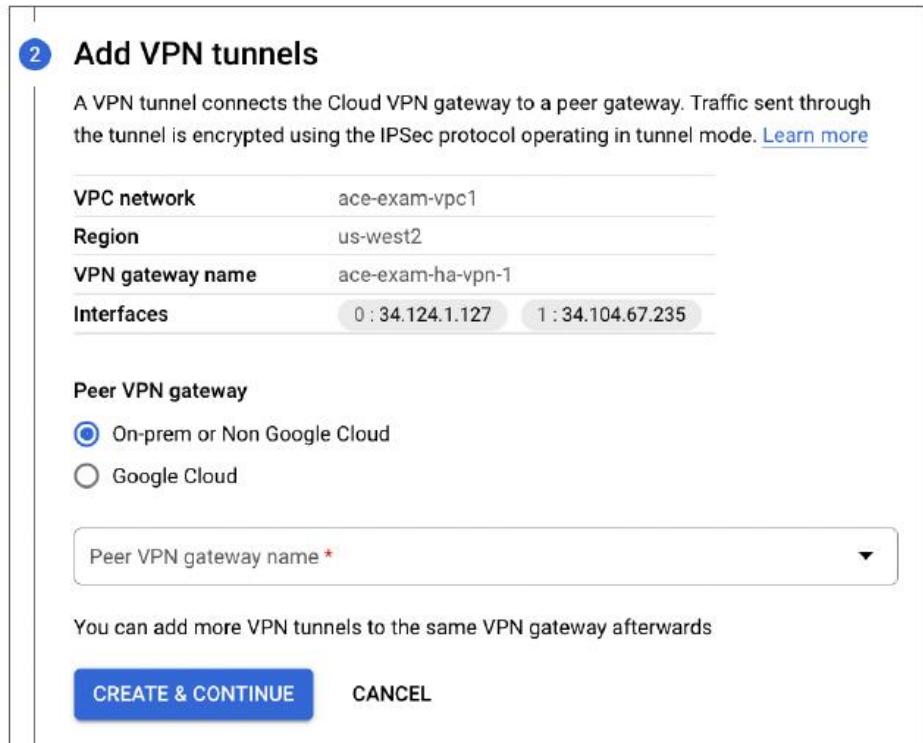
- VPN gateway name ***: ace-exam-ha-vpn-1 (with a question mark icon)
- Network ***: ace-exam-vpc1 (with a question mark icon)
- Region ***: us-west2 (Los Angeles) (with a question mark icon)
- VPN gateway public IP address** (with a question mark icon): Two IP addresses will be automatically allocated for each of your gateway interfaces.
- VPN tunnel inner IP stack type**: The IP stack type will apply to all the tunnels associated with this VPN gateway.
 - IPv4 (single-stack)
 - IPv4 and IPv6 (dual-stack) (with a question mark icon)

At the bottom are 'CREATE & CONTINUE' and 'CANCEL' buttons. To the right of the form, a vertical sidebar lists the steps:

- 2 Add VPN tunnels
- 3 Configure BGP sessions
- 4 Summary and reminder

Em seguida, você adicionará túneis VPN (veja a Figura 14.17). Túneis conectam o gateway VPN a um gateway de pares que existe em um local físico, em outra nuvem ou no Google Cloud.

FIGURE 14.17 Configuring tunnels in an HA VPN



Na seção de Túneis, você configura o outro ponto de extremidade da rede na VPN. Você especifica um nome, descrição e endereço IP do gateway VPN na sua rede. Você terá a opção de escolher um gateway VPN de pares existente ou criar um. Se escolher criar um gateway VPN de pares, você precisará fornecer um nome para ele; especificar 1, 2 ou 4 interfaces; e fornecer o endereço IP externo.

Mundo Real

Análise na Nuvem

A ciência de dados e a análise de dados são cada vez mais importantes para os negócios. Para derivar percepções dessas práticas, você precisa tanto dos dados quanto das ferramentas. Dados sobre clientes, vendas e outros tipos de transações geralmente são armazenados em um banco de dados no centro de dados de uma empresa. As ferramentas que os analistas desejam usar, como Spark e serviços de aprendizado de máquina, estão prontamente disponíveis na nuvem. Muitas organizações têm práticas de segurança para proteger dados e não permitiriam que um analista, por exemplo, baixasse alguns dados e depois os copiasse por uma conexão de Internet insegura para a nuvem. Em vez disso, engenheiros de rede e nuvem criariam uma VPN entre o centro de dados da empresa e o Google Cloud. Isso garantiria que o tráfego de rede entre o centro de dados e a nuvem seja criptografado. Os analistas obtêm acesso aos dados e ferramentas de que precisam, e os profissionais de segurança da informação na organização são capazes de proteger a confidencialidade e integridade dos dados.

Criando uma Rede Privada Virtual Usando gcloud

Para criar uma VPN na linha de comando, você pode usar estes três comandos:

- gcloud compute target-vpn-gateways
- gcloud compute forwarding-rules
- gcloud compute vpn-tunnels

O formato do comando gcloud compute target-vpn-gateways para criar uma VPN Clássica é o seguinte:

```
gcloud compute vpn-tunnels create NOME --peer-address=ENDERECO_PEER --  
shared-secret=SEGREDO_COMPARTILHADO --target-vpn-  
gateway=GATEWAY_VPN_ALVO
```

NOME é o nome do túnel. ENDERECO_PEER é o endereço IPv4 do ponto de extremidade remoto do túnel. SEGREDO_COMPARTILHADO é uma string secreta. GATEWAY_VPN_ALVO é uma referência ao gateway VPN alvo.

Ao criar uma VPN HA, você precisará especificar também o parâmetro --peer-gcp-gateway ou --peer-external-gateway.

O formato do gcloud compute forwarding-rules é o seguinte:

```
gcloud compute forwarding-rules create NOME --target=ESPECIFICACAO_ALVO  
NOME é o nome da regra de encaminhamento. ESPECIFICACAO_ALVO é um dos vários tipos de alvos, incluindo target-instance, target-http-proxy e --target-vpn-gateway.
```

Para detalhes adicionais, veja a documentação em <https://cloud.google.com/sdk/gcloud/reference/compute/forwarding-rules/create>.

O formato do comando gcloud compute vpn-tunnels é o seguinte:

```
gcloud compute vpn-tunnels create NOME --peer-address=ENDERECO_PEER --  
shared-secret=SEGREDO_COMPARTILHADO --target-vpn-  
gateway=GATEWAY_VPN_ALVO
```

NOME é o nome do túnel VPN, ENDERECO_PEER é o endereço IPv4 do túnel remoto, SEGREDO_COMPARTILHADO é uma string secreta e GATEWAY_VPN_ALVO é uma referência a um gateway VPN.

Resumo

Este capítulo revisou como criar VPCs e VPNs. VPCs definem redes no Google Cloud para vincular seus recursos do Google Cloud. VPNs no Google Cloud são usadas para vincular suas redes do Google Cloud às suas redes internas. Discutimos como criar VPCs, VPCs compartilhadas e sub-redes, e descrevemos a notação CIDR. Você também aprendeu como configurar VMs com conexões de rede personalizadas. Em seguida, revisamos regras de firewall e como criá-las. O capítulo concluiu discutindo os passos necessários para criar uma VPN.

Essenciais para o Exame

Saiba que VPCs são centros de dados lógicos na nuvem e que VPNs são conexões seguras entre suas sub-redes VPC e sua rede interna. Seus recursos na nuvem estão em uma VPC. VPCs têm sub-redes e regras de roteamento para direcionar o tráfego entre sub-redes. Você controla o fluxo de tráfego usando regras de firewall.

Saiba que VPCs criam sub-redes em cada região quando estão em modo automático. Você pode criar sub-redes adicionais. Cada sub-rede tem uma faixa de endereços IP. Regras de firewall são aplicadas a sub-redes, também chamadas de redes. Roteadores podem ser configurados para aprender apenas rotas regionais ou rotas globais.

Entenda como ler e calcular a notação CIDR. A notação CIDR representa uma máscara de sub-rede e o tamanho do endereço IP disponível na faixa de IP. Quanto menor o tamanho da máscara de sub-rede, que é o número após a barra em um bloco CIDR, mais endereços IP estão disponíveis.

Saiba que VPCs podem ser criadas usando comandos gcloud. Uma VPC pode ser criada com gcloud compute networks create. Uma VPC compartilhada pode ser criada usando gcloud beta compute shared-vpc. VPCs compartilhadas podem ser compartilhadas no nível de rede ou de pasta. Você precisará vincular políticas de gerenciamento de identidade e acesso (IAM) no nível organizacional ou de pasta para habilitar funções de Administração de VPC Compartilhada. Emparelhamento de VPC pode ser usado para conectividade interprojetos.

Entenda que você pode adicionar interfaces de rede a uma VM. Você pode configurar essas interfaces para usar uma sub-rede específica. Você pode atribuir endereços IP efêmeros ou estáticos.

Saiba que regras de firewall controlam o fluxo de tráfego de rede. Regras de firewall consistem em direção, prioridade, ação, alvo, origem/destino, protocolos e porta, e status de aplicação. Regras de firewall são aplicadas a uma sub-rede.

Saiba como criar uma VPN com o Cloud Console. VPNs roteiam o tráfego entre seus recursos na nuvem e sua rede interna. VPNs incluem gateways, regras de encaminhamento e túneis. Tanto VPNs Clássicas quanto de Alta Disponibilidade (HA) estão disponíveis.

Questões de Revisão

Você pode encontrar as respostas no Apêndice.

1. Que tipo de recurso são as nuvens privadas virtuais no Google Cloud?
 - A. Zonal
 - B. Regional
 - C. Super-regional
 - D. Global
2. Foi-lhe incumbida a tarefa de definir faixas CIDR para usar com um projeto. O projeto inclui duas VPCs com várias sub-redes em cada VPC. Quantas faixas CIDR você precisará definir?
 - A. Uma para cada VPC
 - B. Uma para cada sub-rede
 - C. Uma para cada região
 - D. Uma para cada zona
3. O departamento jurídico precisa isolar seus recursos em sua própria VPC. Você deseja que a rede forneça roteamento para qualquer outro serviço disponível na rede global. A rede VPC não aprendeu rotas globais. Que parâmetro pode ter sido esquecido ao criar as sub-redes da VPC?
 - A. Política de servidor DNS
 - B. Roteamento dinâmico
 - C. Política de roteamento estático
 - D. Política de roteamento sistêmico
4. O comando usado para criar uma VPC a partir da linha de comando é:
 - A. gcloud compute networks create
 - B. gcloud networks vpc create
 - C. gsutil networks vpc create
 - D. gcloud compute create networks
5. Você criou várias sub-redes. A maioria delas está enviando logs para o Cloud Logging. Uma sub-rede não está enviando logs. Qual opção pode ter sido configurada incorretamente ao criar a sub-rede que não está encaminhando logs?
 - A. Logs de Fluxo
 - B. Acesso IP Privado
 - C. Cloud Logging

- D. Máscara de sub-rede de comprimento variável
6. Em quais níveis da hierarquia de recursos uma VPC compartilhada pode ser criada?
- A. Pastas e recursos
 - B. Organizações e projetos
 - C. Organizações e pastas
 - D. Pastas e sub-redes
7. Você está usando o Cloud Console para criar uma VM que você deseja que exista em uma sub-rede personalizada que você acabou de criar. Em qual seção da página Criar Instância você usaria para especificar a sub-rede personalizada?
- A. Aba de Rede da seção Gerenciamento, Segurança, Discos, Rede, Locação Exclusiva
 - B. Aba de Gerenciamento da seção Gerenciamento, Segurança, Discos, Rede, Locação Exclusiva
 - C. Aba de Locação Exclusiva de Gerenciamento, Segurança, Discos, Rede, Locação Exclusiva
 - D. Aba de Locação Exclusiva de Gerenciamento, Segurança, Discos, Rede
8. Você quer implementar a comunicação interprojetos entre VPCs. Qual recurso das VPCs você usaria para implementar isso?
- A. Emparelhamento de rede VPC
 - B. Emparelhamento interprojetos
 - C. VPN
 - D. Interconexão
9. Você quer limitar o tráfego para um conjunto de instâncias. Você decide definir uma tag de rede específica em cada instância. Qual parte de uma regra de firewall pode referenciar a tag de rede para determinar o conjunto de instâncias afetadas pela regra?
- A. Ação
 - B. Alvo
 - C. Prioridade
 - D. Direção
10. Qual parte de uma regra de firewall determina se a regra se aplica ao tráfego de entrada ou de saída?

- A. Ação
 - B. Alvo
 - C. Prioridade
 - D. Direção
11. Você quer definir uma faixa CIDR que se aplica a todos os endereços de destino. Qual endereço IP você especificaria?
- A. 0.0.0.0/0
 - B. 10.0.0.0/8
 - C. 172.16.0.0/12
 - D. 192.168.0.0/16
12. Você está usando o gcloud para criar uma regra de firewall. Qual comando você usaria?
- A. gcloud network firewall-rules create
 - B. gcloud compute firewall-rules create
 - C. gcloud network rules create
 - D. gcloud compute rules create
13. Você está usando o gcloud para criar uma regra de firewall. Qual parâmetro você usaria para especificar a sub-rede à qual ela deve se aplicar?
- A. —subnet
 - B. —network
 - C. —destination
 - D. —source-ranges
14. Uma equipe de desenvolvimento de aplicativos está implantando um conjunto de pontos finais de serviço especializados e quer limitar o tráfego de modo que apenas o tráfego indo para um dos pontos finais seja permitido pelas regras de firewall. Os pontos finais de serviço aceitarão qualquer tráfego UDP, e cada ponto final usará uma porta na faixa de 20000–30000. Qual dos seguintes comandos você usaria?
- A. gcloud compute firewall-rules create fwr1 --allow=udp:20000-30000 --direction=ingress
 - B. gcloud network firewall-rules create fwr1 --allow=udp:20000-30000 --direction=ingress
 - C. gcloud compute firewall-rules create fwr1 --allow=udp
 - D. gcloud compute firewall-rules create fwr1 --direction=ingress

15. Você tem uma regra para permitir tráfego de entrada para uma VM. Você quer que ela se aplique apenas se não houver outra regra que negaria esse tráfego. Qual prioridade você daria a essa regra?
- A. 0
 - B. 1
 - C. 1000
 - D. 65535
16. Você quer criar uma VPN usando o Cloud Console. Em qual seção do Cloud Console você deve usar?
- A. Compute Engine
 - B. App Engine
 - C. Hybrid Connectivity
 - D. IAM & Admin
17. Sua empresa precisa garantir que tenham pelo menos 99,99% de SLA de disponibilidade para redes entre redes locais e uma VPC no Google Cloud. O que você deveria usar para garantir esse nível de disponibilidade?
- A. Classic VPN
 - B. HA VPN
 - C. Shared VPC
 - D. Emparelhamento de rede VPC
18. Você quer que o roteador em um túnel que você está criando aprenda rotas de todas as regiões do Google Cloud na rede. Qual recurso de roteamento do Google Cloud você habilitaria?
- A. Global dynamic routing
 - B. Regional routing
 - C. VPC
 - D. Firewall rules
19. Qual comando gcloud você usaria para criar túneis para uma VPN?
- A. gcloud network vpn-tunnels create
 - B. gcloud compute vpn-tunnels create
 - C. gcloud network create vpn-tunnels
 - D. gcloud compute create vpn-tunnels
20. Você está usando o gcloud para criar uma VPN. Quais comandos você usaria?

- A. gcloud compute target-vpn-gateways only
- B. gcloud compute forwarding-rule e gcloud compute target-vpn-gateways only
- C. gcloud compute vpn-tunnels only
- D. gcloud compute forwarding-rule, gcloud compute target-vpn-gateways, e gcloud compute vpn-tunnels

Capítulo 15

Redes na Nuvem: DNS, Balanceamento de Carga, Acesso Privado Google e Endereçamento IP

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO DE ENGENHEIRO ASSOCIADO DA GOOGLE CLOUD:

- ✓✓ 2.4 Planejamento e configuração de recursos de rede
- ✓✓ 3.5 Implantação e implementação de recursos de rede
- ✓✓ 4.5 Gerenciamento de recursos de rede

Este capítulo continua o foco em redes, especificamente configurando o Sistema de Nomes de Domínio (DNS), balanceamento de carga, Acesso Privado Google e gerenciamento de endereços IP. O Cloud DNS é um serviço gerenciado que fornece serviços de nomeação de domínio autoritativos. Ele é projetado para alta disponibilidade, baixa latência e escalabilidade. Os serviços de balanceamento de carga no Google Cloud oferecem vários tipos de平衡adores de carga para atender a uma gama de necessidades. Neste capítulo, você verá como os平衡adores de carga HTTP(S), Proxy SSL, Proxy TCP, TCP/UDP de Rede e Interno TCP/UDP de Rede diferem e quando usar cada um. Engenheiros de nuvem também devem estar familiarizados com o gerenciamento de endereços IP, em particular gerenciando blocos de roteamento interdomínio sem classe (CIDR) e entendendo como reservar endereços IP. Este capítulo, em combinação com o Capítulo 14, “Redes na Nuvem: Nuvens Privadas Virtuais e Redes Privadas Virtuais”, fornece uma visão geral dos tópicos de rede cobertos no exame de Engenheiro Associado de Nuvem.

Configurando o Cloud DNS

O Cloud DNS é um serviço do Google que fornece resolução de nomes de domínio. No nível mais básico, os serviços DNS mapeiam nomes de domínio, como example.com, para endereços IP, como 35.20.24.107. Uma zona gerenciada contém registros DNS associados a um sufixo de nome DNS, como aceexamdns1.com. Registros DNS contêm detalhes específicos sobre uma zona. Por exemplo, um registro A mapeia um nome de host para endereços IP em IPv4. Registros AAAA são usados em IPv6 para mapear nomes para endereços IPv6. Registros CNAME mapeiam um alias para o nome canônico do domínio. Nesta seção, você aprenderá como configurar serviços DNS no Google Cloud, que consiste em criar zonas e adicionar registros.

Criando Zonas Gerenciadas DNS Usando o Cloud Console

Para criar uma zona gerenciada usando o Cloud Console, navegue até a seção Serviços de Rede do console. Clique em Cloud DNS para acessar a página mostrada na Figura 15.1.

Clique em Criar Zona para acessar a página mostrada na Figura 15.2.

Primeiro, selecione um tipo de zona, que pode ser Pública ou Privada. Em seguida, especifique um nome de zona, que deve ser único dentro do projeto.

FIGURE 15.1 Network Services Cloud DNS page

The screenshot shows the Google Cloud Network Services Cloud DNS page. On the left, there's a sidebar with icons for various services: Load balancing, Cloud DNS (which is selected and highlighted in blue), Cloud CDN, Cloud NAT, Traffic Director, Service Directory, Cloud Domains, Private Service Connect, Marketplace, and Release Notes. The main content area has a header with 'Cloud DNS' and buttons for 'CREATE ZONE' and 'REFRESH'. Below the header are three tabs: 'ZONES' (which is underlined in blue), 'DNS SERVER POLICIES', and 'RESPONSE POLICY ZONES'. A descriptive text box says: 'DNS zones let you define your namespace. You can create public or private zones.' It includes a link to 'Learn more'. In the center, there's a graphic of a globe with colored dots (yellow, green, blue) representing network nodes. Below the graphic, it says 'No DNS Zones' and 'Currently you don't have any DNS zones. Get started by creating one.' At the bottom right is a large blue 'CREATE ZONE' button.

Zonas públicas são acessíveis pela Internet. Essas zonas fornecem servidores de nomes que respondem a consultas de qualquer fonte. Zonas privadas fornecem serviços de nome para seus recursos do Google Cloud, como máquinas virtuais (VMs) e平衡adores de carga. Zonas privadas respondem apenas a consultas originadas de recursos no mesmo projeto que a zona.

No formulário, forneça um nome de zona e descrição. Especifique o nome DNS, que deve ser o sufixo de um nome DNS, como aceexamdns1.com.

Você pode habilitar o DNSSEC, que é a segurança DNS. Ele fornece autenticação forte de clientes que se comunicam com serviços DNS. O DNSSEC é projetado para prevenir spoofing (um cliente parecendo ser algum outro cliente) e envenenamento de cache (um cliente enviando informações incorretas para atualizar o servidor DNS).

Se você escolher criar uma zona privada, terá a opção de escolher configurações que fornecem configurações adicionais para uma zona privada, conforme mostrado na Figura 15.3.

Além dos parâmetros definidos para uma zona pública, você precisará especificar as redes que terão acesso à zona privada.

Depois de criar algumas zonas, a página Cloud DNS listará as zonas, conforme mostrado na Figura 15.4.

FIGURE 15.2 Creating a public DNS zone

The screenshot shows the 'Create a DNS zone' interface. At the top, there's a back arrow and the title 'Create a DNS zone'. Below the title, a descriptive text explains that a DNS zone is a container of DNS records for the same DNS name suffix. It mentions that in Cloud DNS, all records in a managed zone are hosted on the same set of Google-operated authoritative name servers, with a link to 'Learn more'. A note below says 'If you don't have a domain yet, purchase one through [Cloud Domains](#)'. The 'Zone type' section has two options: 'Private' (radio button) and 'Public' (radio button, selected). The 'Zone name' field is a text input with a placeholder 'example-zone-name'. The 'DNS name' field is another text input with a placeholder 'myzone.example.com'. The 'DNSSEC' section has a dropdown menu set to 'Off'. There's a 'Description' text area. Under 'Cloud Logging', there's a dropdown menu set to 'Off'. At the bottom, there are 'CREATE' and 'CANCEL' buttons, and a 'EQUIVALENT COMMAND LINE' dropdown menu.

FIGURE 15.3 Additional configuration options for private DNS zones

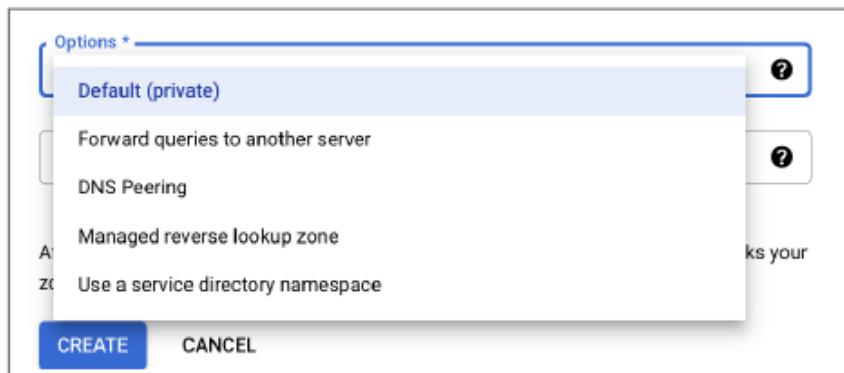


FIGURE 15.4 List of DNS zones

Network services	Cloud DNS			
	CREATE ZONE	REFRESH		
	ZONES	DNS SERVER POLICIES	RESPONSE POLICY ZONES	
DNS zones let you define your namespace. You can create public or private zones. Learn more				
Load balancing				
Cloud DNS				
Cloud CDN				
Cloud NAT				
Traffic Director				
Service Directory				
Cloud Domains				
Private Service Connect				

Clique no nome de uma zona para ver seus detalhes. Conforme mostrado na Figura 15.5, os detalhes da zona incluem uma lista de registros associados à zona. Quando uma zona é criada, registros NS e SOA são adicionados. NS é um registro de servidor de nomes que possui o endereço de um servidor autoritativo que gerencia as informações da zona. SOA é um registro de início de autoridade, que possui informações autoritativas sobre a zona. Você pode adicionar outros registros, como registros A e CNAME.

FIGURE 15.5 List of records in a DNS zone

The screenshot shows a web-based interface for managing a DNS zone. At the top, there are navigation links: 'Zone details' (with a back arrow), 'EDIT', 'ADD RECORD SET', 'ADD NETWORKS', and 'DELETE ZONE'. Below this, the zone name 'ace-exam-dns-zone' is displayed, along with its 'DNS name' (aceexamdns1.com.) and 'Type' (Private). A table titled 'RECORD SETS' lists existing records:

DNS name ↑	Type	TTL (seconds)	Routing policy
aceexamdns1.com.	SOA	21600	Default
aceexamdns1.com.	NS	21600	Default

Below the table are buttons for 'ADD RECORD SET', 'DELETE RECORD SETS', and 'REFRESH'. There is also a 'Filter' option to search for record sets.

Para adicionar um registro A, clique em Adicionar Conjunto de Registros para exibir a página mostrada na Figura 15.6.

FIGURE 15.6 Creating an A record set

The dialog box is titled 'Create record set' and has a back arrow at the top left. It contains the following fields:

- DNS Name:** .aceexamdns1.com. (with a question mark icon)
- Resource Record Type:** A (selected) (with a question mark icon)
- TTL *:** 5 (with a question mark icon)
- TTL Unit:** minutes (with a question mark icon)
- Routing Policy:** Default record type (radio button selected)
- IPv4 Address:** (with a question mark icon)
- IPv4 Address 1 ***: An input field containing '192.0.2.91' with a note 'Example: 192.0.2.91' below it.
- + ADD ITEM**: A button to add more IPv4 addresses.
- CREATE** and **CANCEL**: Buttons at the bottom.

Selecione A como um tipo de registro de recurso e especifique um endereço IPv4 do servidor que mapeia nomes de domínio para endereços IP para esta zona.

Os parâmetros TTL (tempo de vida) e Unidade TTL especificam quanto tempo o registro pode viver em um cache, em outras palavras, o período de tempo em que os

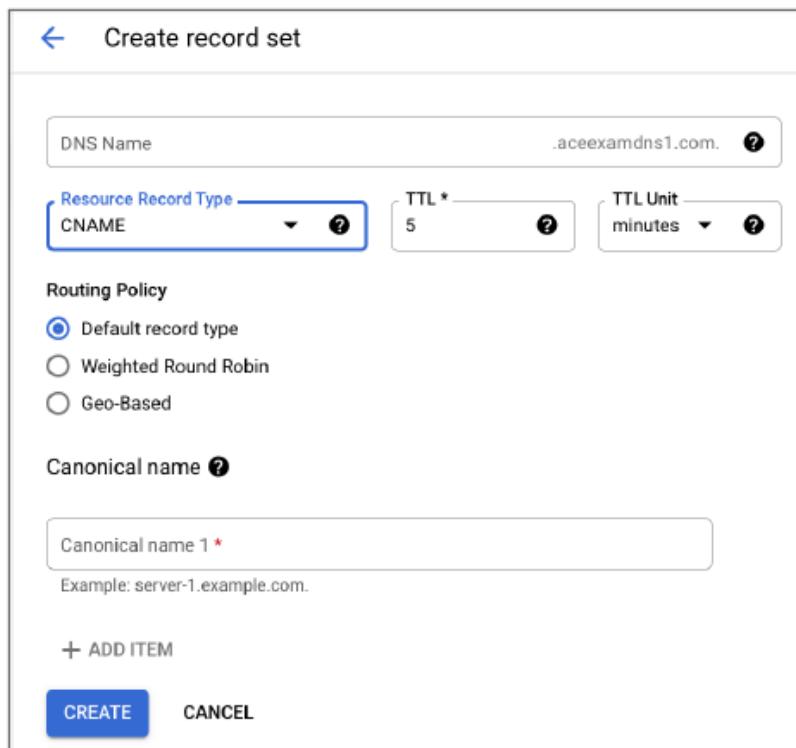
resolvers DNS devem armazenar os dados em cache antes de consultar novamente o valor. Resolvers DNS realizam operações de busca mapeando nomes de domínio para endereços IP. Se você quiser especificar vários endereços IP no registro, clique em Adicionar Item para adicionar outros endereços IP.

Você também pode adicionar registros de nome canônico usando a página Criar Conjunto de Registros. Neste caso, selecione CNAME como o Tipo de Registro de Recurso, conforme mostrado na Figura 15.7.

O registro CNAME leva um nome ou alias de um servidor. Os parâmetros de nome DNS e TTL são os mesmos que no exemplo de registro A.

Além disso, o Encaminhamento DNS agora está disponível, o que permite que suas consultas DNS sejam passadas para um servidor DNS local se você estiver usando o Cloud VPN ou Interconnect..

FIGURE 15.7 Creating a CNAME record



Criando Zonas Gerenciadas DNS Usando gcloud

Para criar zonas DNS e adicionar registros, você usará gcloud dns managed-zones e gcloud dns record-sets transaction.

Para criar uma zona pública gerenciada chamada ace-exam-zone1 com o sufixo DNS aceexamzone.com, use isto:

```
gcloud dns managed-zones create ace-exam-zone1 --description="Uma zona de exemplo" --dns-name=aceexamzone.com.
```

Para tornar isso uma zona privada, você adiciona o parâmetro --visibility definido como private:

```
gcloud dns managed-zones create ace-exam-zone1 --description="Uma zona de exemplo" --dns-name=aceexamzone.com. --visibility=private --networks=default
```

Para adicionar um registro A, você inicia uma transação, adiciona as informações do registro A e então executa a transação.

Transações são iniciadas com gcloud dns record-sets transaction start. Conjuntos de registros são adicionados usando gcloud dns record-sets transaction add, e as transações são concluídas usando gcloud dns record-sets transaction execute. Juntos, os passos são os seguintes:

```
gcloud dns record-sets transaction start --zone=ace-exam-zone1  
gcloud dns record-sets transaction add 192.0.2.91 --name=aceexamzone.com. --ttl=300 -  
-type=A --zone=ace-exam-zone1  
gcloud dns record-sets transaction execute --zone=ace-exam-zone1.
```

Para criar um registro CNAME, você usaria comandos similares:

```
gcloud dns record-sets transaction start --zone=ace-exam-zone1  
gcloud dns record-sets transaction add server1.aceexamzone.com. --  
--name=www2.aceexamzone.com. --ttl=300 --type=CNAME --zone=ace-exam-zone1  
gcloud dns record-sets transaction execute --zone=ace-exam-zone1
```

Configurando Balanceadores de Carga

Balanceadores de carga distribuem a carga de trabalho para servidores executando uma aplicação. Nesta seção, discutiremos os diferentes tipos de balanceadores de carga e como configura-los.

Tipos de Balanceadores de Carga

Balanceadores de carga podem distribuir carga dentro de uma única região ou através de múltiplas regiões. Os vários平衡adores de carga oferecidos pelo Google Cloud são caracterizados por três recursos:

- Balanceamento de carga global versus regional
- Balanceamento de carga externo versus interno
- Tipo de tráfego, como HTTP e TCP

Balanceadores de carga globais são usados quando uma aplicação é distribuída globalmente. Existem quatro balanceadores de carga globais:

- Balanceamento de Carga Externo Global HTTP(S), que balanceia cargas HTTP e HTTPS através de um conjunto de instâncias back-end globalmente em uma camada de serviço de rede Premium.
- Balanceamento de Carga Externo Global HTTP(S) (clássico), que balanceia cargas HTTP e HTTPS através de um conjunto de instâncias back-end

globalmente em rede de camada Premium e regionalmente em rede de camada Padrão.

- Proxy SSL, que termina conexões SSL/TLS, que são conexões de Camada de Soquete Seguro. Este tipo é usado para tráfego não HTTPS.
- Proxy TCP, que termina sessões TCP no balanceador de carga e depois encaminha o tráfego para servidores back-end.

Balanceadores de carga regionais são usados quando os recursos que fornecem uma aplicação estão em uma única região. Os balanceadores de carga regionais são os seguintes:

- Balanceamento de Carga Externo Regional HTTP(S), que balanceia HTTP(S) regionalmente em rede de camada Padrão
- Balanceamento de Carga Interno HTTP(S), que balanceia HTTP(S) regionalmente apenas em rede de camada Premium
- Balanceamento de Carga Interno TCP/UDP, que balanceia TCP/UDP regionalmente apenas em rede de camada Premium
- Balanceamento de Carga Externo de Rede TCP/UDP, que permite o balanceamento de TCP, UDP e outros protocolos regionalmente em rede de camada Padrão ou Premium

Mundo Real

Balanceamento de Carga e Alta Disponibilidade

Aplicações que precisam ser altamente disponíveis devem usar平衡adores de carga para distribuir o tráfego e para monitorar a saúde das VMs no back-end. Uma empresa que oferece acesso à API para dados de clientes precisará considerar como escalar para cima e para baixo em resposta a mudanças na carga e como garantir alta disponibilidade.

A combinação de grupos de instâncias (Capítulo 6, “Gerenciando Máquinas Virtuais”) e平衡adores de carga resolve ambos os problemas. Grupos de instâncias podem gerenciar o autoescalameto, e平衡adores de carga podem realizar verificações de saúde. Se uma VM não estiver funcionando, as verificações de saúde falharão e tirarão a VM com falha da rotação de tráfego. Os usuários da API têm menos probabilidade de receber códigos de resposta falhos quando grupos de instâncias mantêm um número apropriado de VMs ativas e os平衡adores de carga impedem que o tráfego seja roteado para servidores com falhas.

Configurando Balanceadores de Carga Usando o Cloud Console

Para criar um平衡ador de carga no Cloud Console, navegue até a seção Serviços de Rede e selecione Balanceamento de Carga, conforme mostrado na Figura 15.8.

O primeiro passo para criar um平衡ador de carga é decidir sobre o tipo. Neste exemplo, você criará um平衡ador de carga TCP (veja a Figura 15.9).

Depois de selecionar a opção de Balanceamento de Carga TCP, aparecerá a página mostrada na Figura 15.10. Selecione Somente Entre Minhas VMs para balanceamento de carga privado. Este平衡ador de carga será usado em uma única região, e você não descarregará o processamento TCP ou SSL.

FIGURE 15.8 Network Services, Load Balancing section

FIGURE 15.9 Create A Load Balancer options

Você precisará especificar se deseja que o平衡ador de carga lide com tráfego da Internet para suas VMs ou apenas entre VMs na sua rede. Em seguida, especifique se deseja suportar uma única região ou várias regiões. Você também especificará um tipo de back-end, que pode ser Serviço de Back-end, Pool de Destinos ou Instância de Destino. Serviço de Back-end permite especificar como distribuir o tráfego, bem como suporte para escoamento de conexão, verificações de saúde TCP, grupos de instâncias gerenciadas e grupos de failover. Pools de Destinos são instâncias dentro de uma região que são identificadas por uma lista ou URLs que especificam quais VMs podem receber tráfego.

FIGURE 15.10 Creating a TCP balancer

[← Create a load balancer](#)

Please answer a few questions to help us select the right load balancing type for your application

Internet facing or internal only

Do you want to load balance traffic from the Internet to your VMs or only between VMs in your network?

From Internet to my VMs
 Only between my VMs

Multiple regions or single region

Do you want to place the backends for your load balancer in a single region or across multiple regions?

Multiple regions (or not sure yet)
 Single region only

Load Balancer type

Do you want a pass-through load balancer or a proxy load balancer? [?](#)

Pass-through [?](#)
 Proxy [?](#)

CONTINUE

A Figura 15.11 mostra os parâmetros para configurar um back-end, incluindo Tipo de Pilha IP, Verificação de Saúde e Afinidade de Sessão.

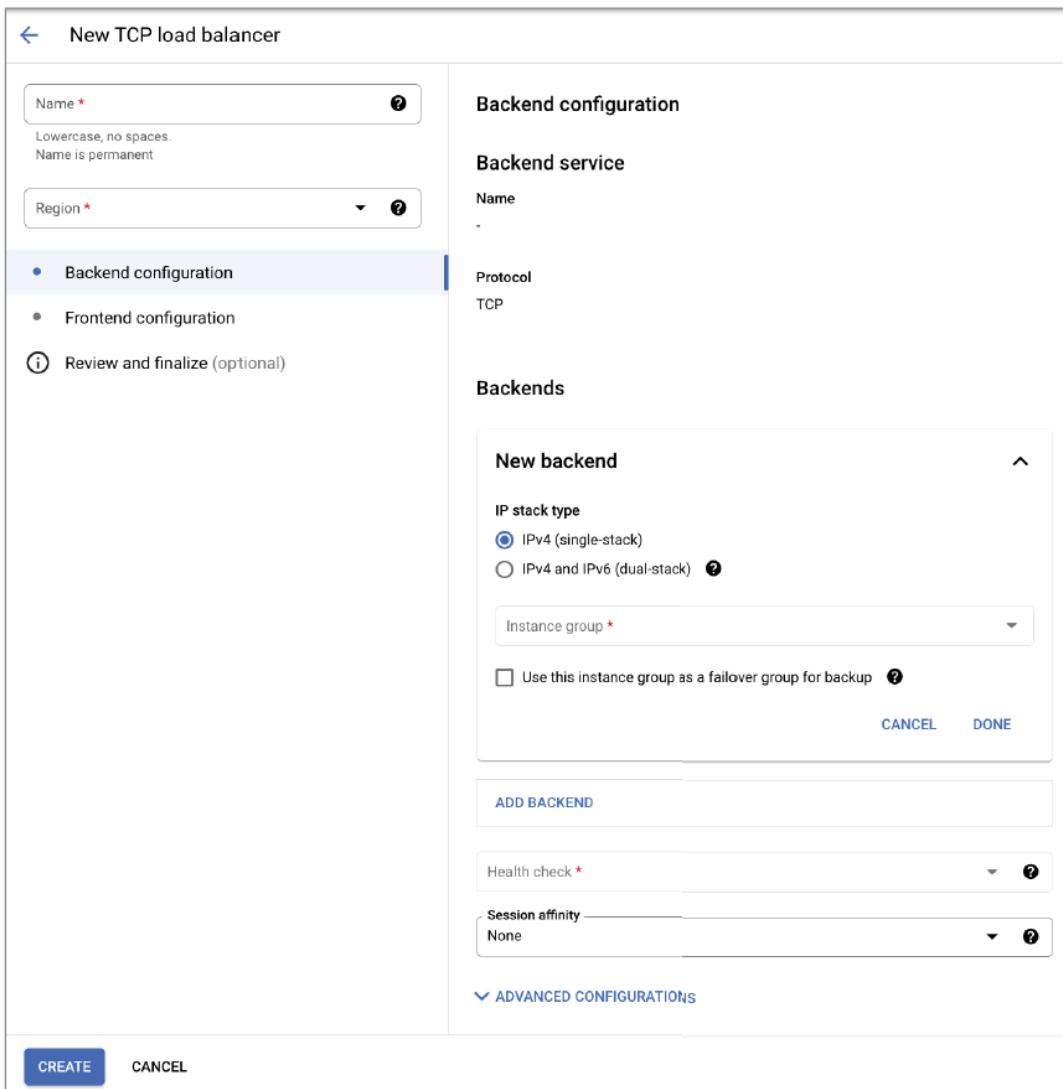
Você pode configurar uma verificação de saúde para o back-end. Isso trará uma página separada, conforme mostrado na Figura 15.12.

Na verificação de saúde, você especifica um nome, um protocolo e uma porta, e um conjunto de critérios de saúde. Neste caso, você verifica os back-ends a cada 5 segundos e aguardará uma resposta por até 5 segundos. Se houver dois períodos consecutivos em que a verificação de saúde falhar, o servidor será considerado não saudável e retirado da rotação do平衡amento de carga.

Em seguida, você configura a frente usando a página na Figura 15.13. Você especifica um nome, sub-rede e uma configuração de IP interno, que neste caso é efêmero (veja “Gerenciando Endereços IP” mais adiante neste capítulo para mais sobre tipos de endereços IP). Você também especifica a porta que terá seu tráfego encaminhado para o back-end. Neste exemplo, você está encaminhando o tráfego na porta 80.

O último passo antes de criar a frente é revisar a configuração e então criar o balanceador de carga.

FIGURE 15.11 Configuring the back end



Configurando Balanceadores de Carga Usando gcloud

Nesta seção, revisaremos as etapas necessárias para criar um balanceador de carga de rede. Estas são boas opções quando você precisa平衡ear protocolos além do HTTP(S).

FIGURE 15.12 Creating a health check

Health Check

Name * ?
Lowercase, no spaces.

Description

Region us-west1 (Oregon) ?

Protocol TCP Port * 80 ?

Proxy protocol NONE ?

Request ? Response ?

Logs
 On Turning on Health check logs can increase costs in Cloud Logging.
 Off

Health criteria
Define how health is determined: how often to check, how long to wait for a response, and how many successful or failed attempts are decisive

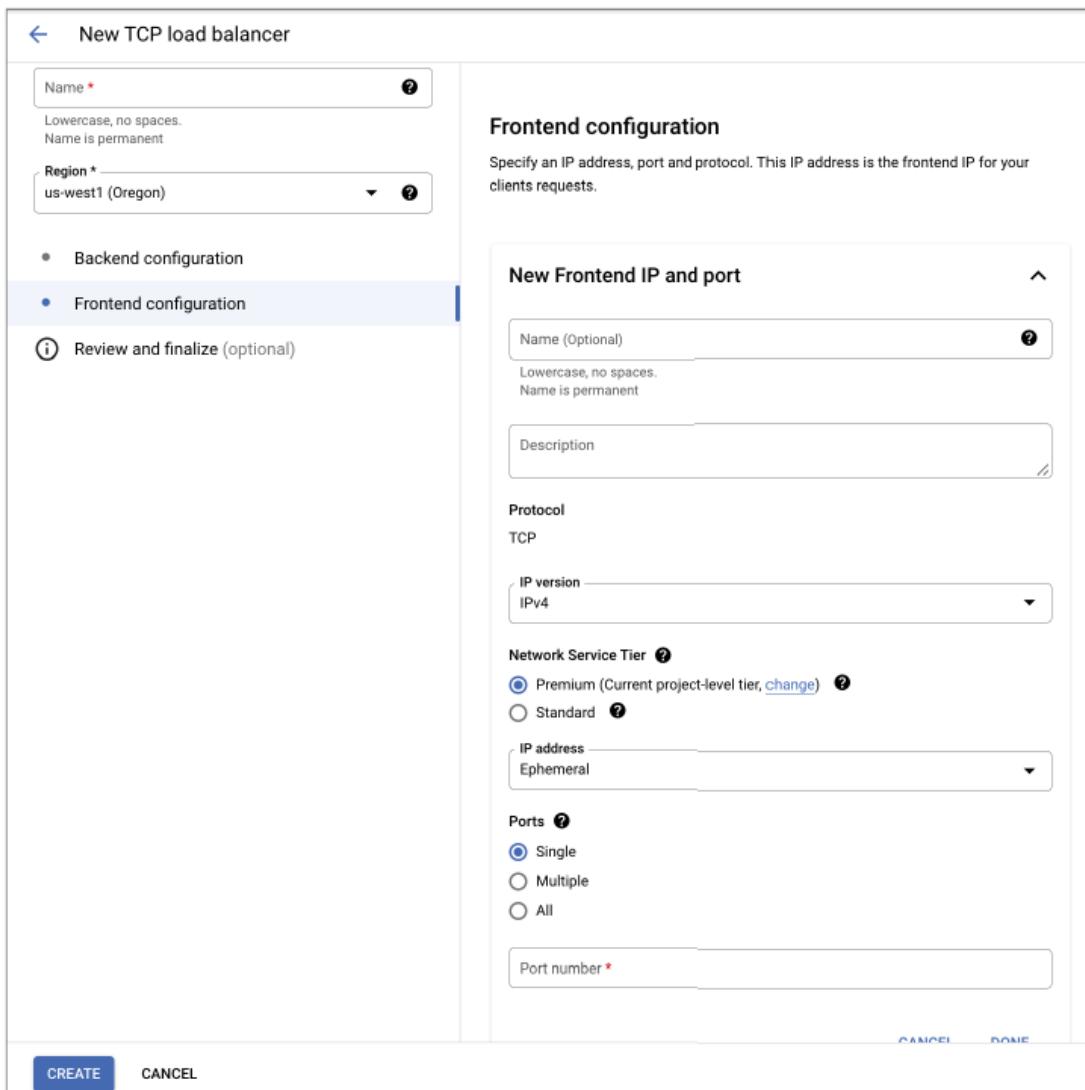
Check interval * 5 seconds ? Timeout * 5 seconds ?

Healthy threshold * 2 consecutive successes ?

Unhealthy threshold * 2 consecutive failures ?

Buttons
SAVE **CANCEL**

FIGURE 15.13 Configuring the front end



O comando `gcloud compute forwarding-rules` é usado para encaminhar tráfego que corresponde a um endereço IP para o平衡ador de carga:

```
gcloud compute forwarding-rules create ace-exam-lb --port=80 --target-pool ace-exam-pool
```

Este comando roteia tráfego para qualquer VM no `ace-exam-pool` para o balanceador de carga chamado `ace-exam-lb`.

Pools de destino são criados usando o comando `gcloud compute target-pools create`. Instâncias são adicionadas ao pool de destino usando o comando `gcloud compute target-pools add-instances`. Por exemplo, para adicionar VMs `ig1` e `ig2` ao pool de destino chamado `ace-exam-pool`, use o seguinte comando:

```
gcloud compute target-pools add-instances ace-exam-pool --instances ig1,ig2
```

Acesso Privado Google

VMs executando em uma VPC podem usar endereços IP externos para se conectar a APIs do Google e outros serviços. No entanto, se você não quiser atribuir endereços IP externos às VMs, você pode usar uma das opções de Acesso Privado Google.

O Acesso Privado Google é usado para alcançar recursos do Google Cloud sem usar um endereço IP externo. Isso permite conectar-se a APIs do Google através do gateway de rede padrão da VPC. Esta opção suporta a maioria dos serviços e APIs do Google Cloud. Tráfego para APIs do Google de sistemas locais precisa ser enviado para `private.googleapis.com` ou `restricted.googleapis.com`. Além disso, rotas devem estar no lugar para o tráfego fluir de sistemas locais para `private.googleapis.com` ou `restricted.googleapis.com`.

O Acesso a Serviços Privados é usado para acessar uma rede VPC gerenciada pelo Google ou por terceiros através de uma conexão de Emparelhamento VPC. Isso suporta alguns serviços do Google Cloud e serviços de terceiros.

O Conecte de Serviços Privados é usado com recursos do Google Cloud que podem ou não ter endereços IP externos, bem como sistemas locais. Com esta opção, você se conecta a um ponto de extremidade de Conecte de Serviços Privados em sua rede VPC e esse ponto de extremidade encaminhará solicitações para APIs e serviços do Google.

Se você estiver usando o Cloud Run, o App Engine Standard e o Cloud Functions, então você pode usar o Acesso VPC Sem Servidor para alcançar endereços IP privados desses serviços.

Gerenciando Endereços IP

Os tópicos do exame para a certificação de Engenheiro Associado à Cloud especificamente identificam dois tópicos relacionados a endereços IP: expandindo blocos CIDR e reservando endereços IP.

Também é importante entender a diferença entre endereços IP efêmeros e estáticos. Endereços IP estáticos são atribuídos a um projeto até que sejam liberados. Eles são usados se você precisar de um endereço IP fixo para um serviço, como um site. Endereços IP efêmeros existem apenas enquanto o recurso estiver usando o endereço IP, como em uma VM executando uma aplicação acessada apenas por outras VMs no mesmo projeto. Se você deletar ou parar uma VM, endereços efêmeros são liberados.

Expandindo Blocos CIDR

Blocos CIDR definem uma faixa de endereços IP disponíveis para uso em uma sub-rede. Se você precisar aumentar o número de endereços disponíveis — por exemplo, se você precisar expandir o tamanho de clusters executando em uma sub-rede — você pode usar o comando `gcloud compute networks subnets expand-ip-range`. Ele recebe o nome da sub-rede e um novo comprimento de prefixo. O comprimento do prefixo determina o tamanho da máscara de rede.

Por exemplo, para aumentar o número de endereços em `ace-exam-subnet1` para 65.536, você define o comprimento do prefixo para 16:

```
gcloud compute networks subnets expand-ip-range ace-exam-subnet1 --prefix-length 16
```

Isso pressupõe que o comprimento do prefixo era maior que 16 antes de emitir este comando. O comando expand-ip-range é usado apenas para aumentar o número de endereços. Você não pode diminuí-los, embora. Você teria que recriar a sub-rede com um número menor de endereços.

Reservando Endereços IP

Endereços IP externos estáticos podem ser reservados usando o Cloud Console ou a linha de comando. Para reservar um endereço IP estático usando o Cloud Console, navegue até a seção Virtual Private Cloud (VPC) do console e selecione Endereços IP. Isso exibirá uma página como a mostrada na Figura 15.14.

FIGURE 15.14 VPC Network IP Address page

Name	IP address	Access type	Region	Type	Version
default-ip-range	10.87.224.0	Internal		Static	IPv4

Clique em Reservar Endereço IP Estático Externo para exibir a página mostrada na Figura 15.15, onde você pode reservar um endereço IP.

Ao reservar um endereço IP, você precisará especificar um nome e uma descrição opcional. Você pode ter a opção de usar a camada de serviço Standard de menor custo para redes, que usa a Internet para alguma transferência de dados. A camada Premium roteia todo o tráfego pela rede global do Google. Você também precisará determinar se o endereço é em IPv4 ou IPv6 e se é regional ou global. Você pode anexar o endereço IP estático a um recurso como parte do processo de reserva, ou você pode mantê-lo não anexado.

FIGURE 15.15 Reserving a static IP address

The screenshot shows the 'Reserve a static address' interface. At the top, there's a back arrow and the title 'Reserve a static address'. Below that is a 'Name *' field with a question mark icon and a note: 'Lowercase letters, numbers, hyphens allowed'. There's also a 'Description' field with a text input area and a question mark icon. The next section is 'Network Service Tier' with two options: 'Premium (Current project-level tier, [change](#))' (selected) and 'Standard'. Below that is 'IP version' with 'IPv4' selected. Under 'Type', 'Regional' is selected. A note says '(to be used with Global forwarding rules [Learn more](#))'. The 'Region' dropdown is set to 'us-central1 (Iowa)'. The 'Attached to' dropdown is set to 'None'. A note below it says 'Some of the instances may be disabled due to the "External IPs for VM instances" organization policy. [Learn more](#)'. A warning message in a box states: '⚠ Static IP addresses not attached to an instance or load balancer are billed at a higher hourly rate [Pricing details](#)'. At the bottom are 'RESERVE' and 'CANCEL' buttons.

Endereços reservados permanecem anexados a uma VM mesmo quando ela não está em uso e permanecem anexados até serem liberados. Isso é diferente dos endereços efêmeros, que são liberados automaticamente quando uma VM é desligada.

Para reservar um endereço IP usando a linha de comando, utilize o comando gcloud:

```
gcloud compute addresses create ace-exam-reserved-static1 --region=us-west2 --network-tier=PREMIUM
```

Por exemplo, para criar um endereço IP estático na região us-west2, que utiliza a camada Premium, use este comando. Isso garante que o endereço IP estático reservado estará disponível para ser associado a recursos específicos no seu projeto do Google Cloud, fornecendo um ponto de acesso consistente e de longo prazo para serviços cruciais, como um site ou aplicação web, mesmo quando as VMs são reiniciadas ou redesenhadas.

Resumo

O exame de Engenheiro Associado à Cloud pode testar seu conhecimento sobre Cloud DNS, balanceamento de carga e gerenciamento de endereços IP. O Cloud DNS é um serviço de nome autoritativo para mapear nomes de domínio para endereços IP. Você pode configurar zonas DNS públicas ou privadas. Você também precisará estar familiarizado com平衡amento de carga e os diferentes tipos de平衡adores de carga. Alguns平衡adores de carga são regionais, e alguns são globais. Alguns são para uso interno apenas, e outros suportam fontes externas de tráfego. O capítulo também revisou como expandir o número de endereços disponíveis em uma sub-rede e discutiu como reservar endereços IP.

Essenciais para o Exame

Entenda que o Cloud DNS é usado para mapear nomes de domínio para endereços IP. Se você deseja suportar consultas da Internet, use uma zona DNS pública. Use uma zona DNS privada apenas se você quiser aceitar consultas de recursos no seu projeto.

Saiba que entradas DNS, como example.com, podem ter vários registros associados a elas. O registro A especifica o endereço de um resolvelor DNS que mapeia nomes de domínio para endereços IP. Registros CNAME armazenam o nome canônico do domínio

Saiba como os平衡adores de carga são diferenciados. Balanceadores de carga são diferenciados com base em平衡amento de carga global versus regional,平衡amento de carga externo versus interno e os protocolos suportados. Balanceadores globais distribuem carga através de regiões, enquanto平衡adores de carga regionais funcionam dentro de uma região. Balanceadores de carga internos balanceiam tráfego apenas de dentro do Google Cloud, não de fontes externas. Alguns平衡adores de carga são específicos de protocolo, como平衡adores de carga HTTP e SSL.

Saiba os tipos de平衡adores de carga e quando devem ser usados. HTTP(S), Proxy SSL, Proxy TCP e TCP/UDP. Balanceadores de carga distribuem carga regional ou globalmente. Balanceadores de carga internos distribuem carga de tráfego interno. Balanceadores de carga externos distribuem carga de tráfego externo.

- HTTP(S) balanceia carga HTTP e HTTPS.
- Proxy SSL termina conexões SSL/TLS.
- Proxy TCP termina sessões TCP.
- TCP/UDP balanceia tráfego TCP/UDP em redes privadas hospedando VMs internas.

Entenda que configurar um平衡ador de carga pode exigir a configuração tanto da frente quanto do back-end. O平衡ador de carga de rede pode ser configurado especificando uma regra de encaminhamento que roteia tráfego para o平衡ador de carga para VMs no pool de destino.

Conheça as opções de Acesso Privado Google. Acesso Privado Google é usado para acesso privado à maioria dos serviços do Google Cloud, enquanto Acesso a Serviços Privados é usado com serviços de terceiros e alguns serviços do Google Cloud. Conecte de Serviços Privados usa um ponto de extremidade VPC para encaminhar tráfego para

serviços do Google Cloud. Acesso VPC Sem Servidor permite que o Cloud Run, Cloud Functions e App Engine Standard alcancem VMs com endereços privados.

Saiba como aumentar o número de endereços IP em uma sub-rede. Use o comando `gcloud compute networks subnets expand-ip-range` para aumentar os endereços IP em uma sub-rede. O número de endereços só pode aumentar. O comando `expand-ip-range` não pode ser usado para diminuir o número de endereços.

Saiba como reservar um endereço IP usando o console e o comando `gcloud compute addresses create`. Endereços IP reservados continuam disponíveis para o seu projeto mesmo se não estiverem anexados a um recurso. Conheça a diferença entre os serviços de rede de camada Premium e Standard.

Questões de Revisão

Você pode encontrar as respostas no Apêndice.

1. Qual tipo de registro é usado para especificar o endereço IPv4 de um domínio?
 - A. AAAA
 - B. A
 - C. NS
 - D. SOA
2. O CEO da sua startup acabou de ler um relatório de notícias sobre uma empresa que foi atacada por algo chamado envenenamento de cache. O CEO quer implementar medidas de segurança adicionais para reduzir o risco de spoofing de DNS e envenenamento de cache. O que você recomendaria?
 - A. Usar DNSSEC
 - B. Adicionar registros SOA
 - C. Adicionar registros CNAME
 - D. Deletar registros CNAME
3. O que os parâmetros TTL especificam em um registro DNS?
 - A. Tempo que um registro pode existir em um cache antes que deva ser consultado novamente
 - B. Tempo que um cliente tem para responder a uma solicitação de informações de DNS
 - C. Tempo permitido para criar um registro CNAME
 - D. Tempo antes de um humano ter que verificar manualmente as informações no registro DNS
4. Qual comando é usado para criar uma zona DNS na linha de comando?
 - A. gsutil dns managed-zones create
 - B. gcloud dns managed-zones create
 - C. gcloud managed-zones create
 - D. gcloud create dns managed zones
5. Qual parâmetro é usado para tornar uma zona DNS privada?
 - A. --private
 - B. --visibility=private

C. --private=true

D. --status=private

6. Quais平衡adores de carga fornecem balanceamento de carga global?

A. Global External HTTP(S) Load Balancing e Global External HTTP(S) Load Balancing (clássico) apenas

B. SSL Proxy e TCP Proxy apenas

C. Global External HTTP(S) Load Balancing, Global External HTTP(S) Load Balancing (clássico), SSL Proxy, e TCP Proxy

D. Internal TCP/UDP, HTTP(S), SSL Proxy, e TCP Proxy

7. Qual balanceador de carga regional balanceia HTTP(S) regionalmente apenas em rede de camada Premium?

A. Global External HTTP(S) Load Balancing

B. SSL Proxy

C. TCP Proxy

D. Internal HTTP(S) Load Balancing

8. Você está configurando um balanceador de carga e quer implementar balanceamento de carga privado. Qual opção você selecionaria?

A. Only Between My VMs

B. Enable Private

C. Disable Public

D. Local Only

9. Quais dois componentes precisam ser configurados ao criar um balanceador de carga TCP Proxy?

A. Frente e regra de encaminhamento

B. Frente e fundo

C. Regra de encaminhamento e fundo apenas

D. Fundo e regra de encaminhamento apenas

10. Uma verificação de saúde é usada para verificar quais recursos?

A. Políticas da organização

B. VMs

C. Buckets de armazenamento

D. Discos persistentes

11. Onde você especifica as portas em um平衡ador de carga TCP Proxy que devem ter seu tráfego encaminhado?
- A. Fundo
 - B. Frente
 - C. Seção Serviços de Rede
 - D. VPC
12. Qual comando é usado para criar um balanceador de carga de rede na linha de comando?
- A. gcloud compute forwarding-rules create
 - B. gcloud network forwarding-rules create
 - C. gcloud compute create forwarding-rules
 - D. gcloud network create forwarding-rules
13. Uma equipe está configurando um serviço web para uso interno. Eles querem usar o mesmo endereço IP por um futuro previsível. Que tipo de endereço IP você atribuiria?
- A. Interno
 - B. Externo
 - C. Estático
 - D. Efêmero
14. Você está iniciando uma VM para experimentar uma nova biblioteca de ciência de dados em Python. Você usará SSH para se conectar à VM, usará o interpretador Python interativamente por um tempo e, então, desligará a máquina. Que tipo de endereço IP você atribuiria a esta VM?
- A. Efêmero
 - B. Estático
 - C. Permanente
 - D. IPv8
15. Você criou uma sub-rede chamada sn1 usando 192.168.0.0 com 65.534 endereços. Você percebe que não precisará de tantos endereços e gostaria de reduzir esse número para 254. Qual dos seguintes comandos você usaria?
- A. gcloud compute networks subnets expand-ip-range sn1 --prefix-length=24
 - B. gcloud compute networks subnets expand-ip-range sn1 --prefix-length=-8
 - C. gcloud compute networks subnets expand-ip-range sn1 --size=256
 - D. Não existe comando para reduzir o número de endereços IP disponíveis.

16. Você criou uma sub-rede chamada sn1 usando 192.168.0.0. Você quer que ela tenha 14 endereços. Qual comprimento de prefixo você usaria?
- A. 32
 - B. 28
 - C. 20
 - D. 16
17. Você quer que todo o seu tráfego de rede seja roteado pela rede do Google e não atravesse a Internet pública. Qual nível de serviço de rede você escolheria?
- A. Padrão
 - B. Somente Google
 - C. Premium
 - D. Não-Internet
18. Você tem um site hospedado em uma VM do Compute Engine. Os usuários podem acessar o site usando o nome de domínio que você forneceu. Você realiza alguns trabalhos de manutenção na VM e para o servidor e reinicia. Agora, os usuários não conseguem acessar o site. Nenhuma outra alteração ocorreu na sub-rede. Qual pode ser a causa do problema?
- A. A reinicialização causou uma mudança no registro DNS.
 - B. Você usou um endereço IP efêmero em vez de um estático.
 - C. Você não tem endereços suficientes disponíveis na sua sub-rede.
 - D. Sua sub-rede mudou.
19. Você está implantando um sistema distribuído. Mensagens serão passadas entre VMs do Compute Engine usando um protocolo UDP confiável. Todas as VMs estão na mesma região. Você quer usar o balanceador de carga que melhor se encaixa nestes requisitos. Que tipo de balanceador de carga você usaria?
- A. TCP/UDP Interno
 - B. Proxy TCP
 - C. Proxy SSL
 - D. Balanceamento de Carga Externo Global HTTP(S)
20. Você quer usar o Cloud Console para revisar os registros em uma entrada DNS. Para qual seção do Cloud Console você navegaria?
- A. Compute Engine
 - B. Serviços de Rede

- C. Kubernetes Engine
- D. Conectividade Híbrida

Capítulo 16

Implantando Aplicações com Cloud Marketplace e Cloud Foundation Toolkit

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO DE ENGENHEIRO DE NUVEM ASSOCIADO DO GOOGLE:

- ✓✓ 3.6 Implantar uma solução usando Cloud Marketplace
- ✓✓ 3.7 Implementar recursos via infraestrutura como código

Ao longo deste guia de estudo, você aprendeu como implantar recursos de computação, armazenamento e rede, e agora você voltará sua atenção para a implantação de aplicações. Cloud Marketplace é um serviço do Google Cloud para encontrar e implantar aplicações pré-configuradas que estão prontas para rodar no Google Cloud. Cloud Marketplace permite que os usuários implantem aplicações e os recursos de computação, armazenamento e rede necessários sem ter que configurar esses recursos eles mesmos. Cloud Foundation Toolkit é um conjunto de ferramentas usado para simplificar a implantação de infraestrutura como código.

Implantando uma Solução Usando Cloud Marketplace

Cloud Marketplace é um repositório central de aplicações e conjuntos de dados que podem ser implantados no seu ambiente Google Cloud. Trabalhar com o Cloud Marketplace é um processo de dois passos: buscar por uma solução que atenda às suas necessidades e, em seguida, implantar a solução.

Navegando pelo Cloud Marketplace e Visualizando Soluções

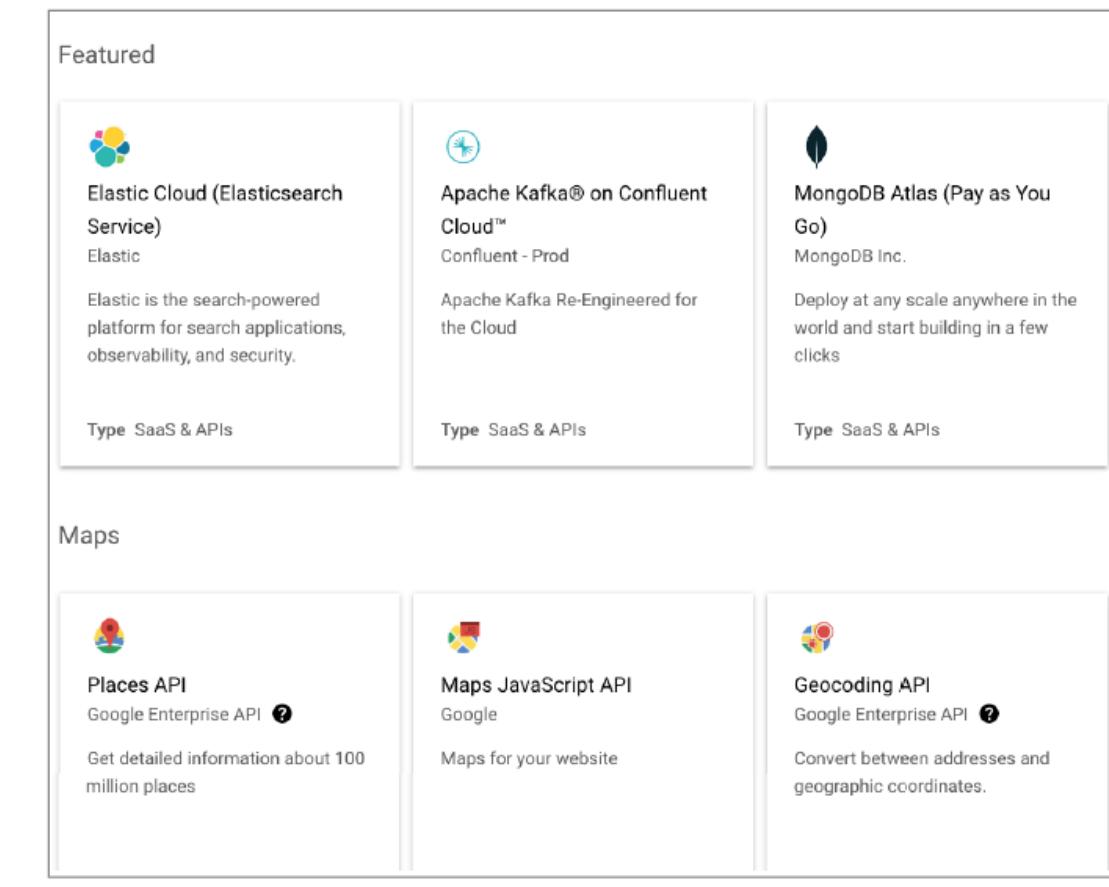
Para visualizar as soluções disponíveis no Cloud Marketplace, navegue até a seção Marketplace. Isso exibirá uma página como a mostrada na Figura 16.1.

A página principal do Cloud Marketplace mostra algumas soluções em destaque. As soluções mostradas na Figura 16.1 incluem Elastic Cloud (Serviço Elasticsearch), Apache Kafka no Confluent Cloud, MongoDB Atlas, bem como APIs para geocodificação e direções.

Você pode buscar ou navegar por filtro para ver a lista de soluções. A Figura 16.2 mostra a lista de categorias de soluções disponíveis.

Você pode restringir o conjunto de soluções exibidas na página principal escolhendo uma categoria específica. Por exemplo, se você filtrar para ver apenas Big Data, verá uma lista de opções, como mostrado na Figura 16.3. Você pode ver uma lista dos sistemas operacionais disponíveis na Figura 16.4.

FIGURE 16.1 Cloud Marketplace main page

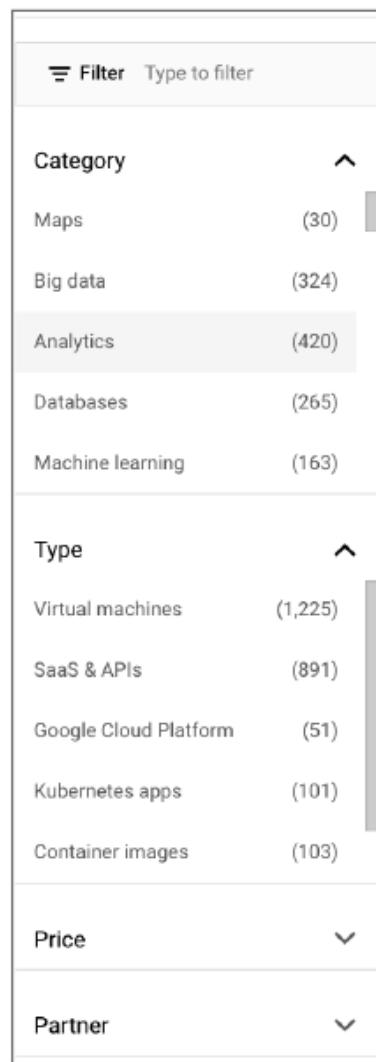


Observe que você pode filtrar ainda mais a lista de sistemas operacionais por tipo de licença. Os tipos de licença são grátis, taxa horária fixa, taxas de uso e traga sua própria licença (BYOL). Os sistemas operacionais gratuitos incluem opções Linux e FreeBSD. Os sistemas operacionais disponíveis por uma taxa incluem Windows e Linux suportado por empresas. Você será cobrado uma taxa baseada no seu uso, e essa cobrança será incluída na sua fatura do Google Cloud. A opção BYOL inclui dois sistemas operacionais Linux suportados que requerem que você tenha uma licença válida para executar o software. Você é responsável por adquirir a licença antes de executar o software.

A Figura 16.5 mostra uma amostra de ferramentas de desenvolvedor disponíveis no Cloud Marketplace. Estas incluem WordPress, Joomla e Alfresco.

Vamos dar uma olhada no tipo de informação fornecida junto com as soluções listadas no Cloud Marketplace. A Figura 16.6 mostra a maior parte das informações disponíveis. Inclui uma visão geral, informações de preços e detalhes sobre o conteúdo do pacote. Há também informações sobre onde a solução será executada dentro do Google Cloud.

FIGURE 16.2 Filtering by category



Informações de Preços (ver Figura 16.7) também são mostradas na página de visão geral. Esses são os custos estimados para rodar a solução, conforme configurado, por um mês, o que inclui os custos de VMs, discos persistentes e quaisquer outros recursos. A estimativa de preço também inclui descontos para uso contínuo de recursos do Google Cloud, que são aplicados à medida que você atinge um limite baseado na quantidade de tempo que um recurso é usado. As últimas seções da página de visão geral do produto fornecem informações e links para documentação e tutoriais, bem como informações de suporte.

FIGURE 16.3 Big Data options available in Cloud Marketplace

The screenshot shows the Google Cloud Marketplace search results for 'Big data'. The search bar at the top has 'Big data' typed in. Below it, the text '324 results' is displayed. The results are presented in a grid format with three columns and four rows.

Category	Type	Description
Type Datasets	Cymbal	About Cymbal: Google Cloud's demo brand Cymbal Group Synthetic datasets across industries showcasing Google Cloud.
	Adobe Analytics 2.0 by Supermetrics	Instantly connect your Adobe Analytics 2.0 data to BigQuery
Type SaaS & APIs	AION	Aion On-Chain Transaction Data cmorq Easy access to on-chain transaction data
	American Community Survey (ACS)	Analytics Hub API Google Exchange data and analytics assets securely and efficiently.
Type Virtual machines	Apache Cassandra Cluster on CentOS Server 8.4	Cloud Infrastructure Services Apache Cassandra, open-source NoSQL database cluster software
	Analytics Hub API Google Exchange data and analytics assets securely and efficiently.	Type SaaS & APIs

Implantando Soluções do Cloud Marketplace

Depois de identificar uma solução que atende às suas necessidades, você pode lançá-la a partir do Cloud Marketplace. Vá para a página de visão geral do produto que você gostaria de lançar, como mostrado na Figura 16.9, e selecione Lançar.

Isso abrirá a página mostrada na Figura 16.10. Você pode ver uma mensagem informando que APIs adicionais devem ser habilitadas para implantar uma solução. Nesse caso, habilite as APIs adicionais, e uma vez que você fizer isso, a página na Figura 16.10 aparecerá.

FIGURE 16.4 Operating systems available in Cloud Marketplace

Operating systems

Featured



Debian 10
Debian
Debian 10 (Buster)

Type Virtual machines



Ubuntu 20.04 LTS (Focal)
Canonical
Ubuntu 20.04 LTS (Focal)

Type Virtual machines



Red Hat Enterprise Linux 8
Red Hat
Red Hat Enterprise Linux 8

Type Virtual machines

177 results



AlmaLinux 8
AlmaLinux
Almal inux 8

Type Virtual machines



Apache Tomcat Server on
Windows 2019
Cloud Infrastructure Services
Tomcat is lightweight, highly
flexible & secure

Type Virtual machines



BlackArch Linux By
Techlatest.net
Out of box Blackarch Pentest VM
with 2800+ preinstalled tools

Type Virtual machines

O conteúdo desta página variará por aplicação, mas muitos parâmetros são comuns entre as soluções. Nesta página, você especifica um nome para a implantação, uma zona e o tipo de máquina. Você pode escolher o tipo e o tamanho do disco persistente.

Neste exemplo, a solução será implantada em um servidor de 2 vCPUs com 8 GB de memória e um disco de inicialização de 10 GB usando discos persistentes padrão. Se desejar, você poderia optar por um disco SSD para o disco de inicialização. Você também pode alterar o tamanho do disco de inicialização.

FIGURE 16.5 Developer tools available in Cloud Marketplace

Developer tools

Featured



WordPress Certified by Bitnami and Automatic Bitnami
Up-to-date, secure, and ready to run.

Type Virtual machines



Joomla! packaged by Bitnami Bitnami
Up-to-date, secure, and ready to run.

Type Virtual machines



Alfresco Community packaged by Bitnami Bitnami
Up-to-date, secure, and ready to run.

Type Virtual machines

431 results



About Cymbal: Google Cloud's demo brand
Cymbal Group
Synthetic datasets across industries showcasing Google Cloud.

Type Datasets



Actifio Global Manager
Actifio
Web scale management for Actifio Sky appliances

Type Virtual machines



Actifio Sky
Actifio
Enterprise Class Backup and Recovery

Type Virtual machines

Na seção de Redes, você pode especificar a rede e a sub-rede para lançar a VM. Você também pode configurar regras de firewall para permitir tráfego HTTP e HTTPS. Além disso, você pode especificar faixas de IP de origem para o tráfego HTTP e HTTPS. Se você expandir a seção de Redes, verá parâmetros adicionais para especificar rede, sub-rede e endereços IP externos. (Ver Figura 16.11.)

FIGURE 16.6 Overview page of a WordPress solution

WordPress Certified by Bitnami and Automatic

Up-to-date, secure, and ready to run.

LUNCH **VIEW PAST DEPLOYMENTS**

OVERVIEW **PRICING** **DOCUMENTATION** **SUPPORT**

Overview

Bitnami, the leaders in application packaging, and Automatic, the experts behind WordPress, have teamed up to offer this official WordPress image on Google Cloud Marketplace.

WordPress is the world's most popular content management platform. Whether it's for an enterprise or small business website, or a personal or corporate blog, content authors can easily create content using its new Gutenberg editor, and developers can extend the base platform with additional features.

For content authors, the Jetpack plugin (enabled by default) offers access to additional professional themes, performance improvements, scanning, site activity, and marketing tools. Other popular plugins like Akismet, All in One SEO Pack, WP Mail and Google Analytics for WordPress also come pre-installed. Optional automatic backup and priority support are available from Automatic.

For developers, this image features the AMP for WordPress plugin. This plugin automatically adds Accelerated Mobile Pages (Google AMP Project) support to deliver a faster, higher-performance and more flexible web experience across distribution platforms. It helps you reduce the operating and development costs of your site by pairing your content to the format required by the destination platform and making the user experience consistent across devices.

This image includes the latest version of WordPress, PHP, Apache, and MySQL. It is secure by default, as all ports except HTTP and HTTPS ports are closed. HTTP/2 and Let's Encrypt auto-configuration are supported.

Additional details

Runs on: Google Compute Engine
Type: [Virtual machines](#), Single VM
Last updated: 7/4/22
Category: [Blog & CMS](#)
Version: 5.0.0-6-r03
Operating System: Debian 11
Package contents: [Apache 2.4.4](#), [Simple Tags 3.0.4](#), [AMP 2.3.0](#), [WordPress Mail SMTP 3.4.0](#), [All-in-One WP Migration 7.6.1.0](#), [All in One SEO Pack 4.2.2-r01](#), [W3 Total Cache 2.2.3-0](#), [Google Analytics Dashboard 8.6.0](#), [Jetpack 9.9.1-0](#), [WordPress Amazon Polly Plugin 4.3.2](#), [WordPress 6.0.0](#), [mod_pagespeed library 1.13.35-2](#), [mod_pagespeed_ap2 library 1.13.35-2](#), [ModSecurity Apache Connector 0.20210819.0](#), [Apache utilities \(APR\) 1.6.1](#), [Apache Portable Runtime \(APR\) 1.7.0](#), [ModSecurity 3.0.7](#), [ModSecurity2.9.5](#), [Apache 2.4.54](#), [Apache PageSpeed Module 1.13.35-2](#), [MariaDB 10.6.8](#), [Composer 2.3.7](#), [PECL APC User Cache 5.1.21](#), [MaxMind DB Reader PHP API 1.11.0](#), [libmemcached 3.2.0](#), [PECL PHP driver for Xdebug 3.1.5](#), [libmemcached 1.6.0](#), [PECL PHP driver for Imagick 3.7.0](#), [PECL PHP driver for MongoDB 1.13.0](#), [IMAP 2007.0.0](#), [PHP 8.0.20](#), [qpress 11.0.6.0](#), [Percona XtraBackup 8.0.28-21](#), [vmraid](#), [queryeting 2.0.3](#), [Varnish 6.6.2](#), [phpMyAdmin 5.2.0](#), [Bridging Diagnostic Tool 0.9.17](#), [wait-for-port 1.0.3](#), [Gonit 0.2.6](#), [MySQL 8.0.29](#), [gesu 1.14.0](#), [Brotli 1.0.9](#), [WP-CJ 2.6.0](#), [Bncert Tool 0.7.4](#), [render-template 1.0.3](#), [ini-file 1.4.3](#)

Add to Service Catalog: [Deployment zip file](#)

FIGURE 16.7 Pricing estimates for the WordPress solution

Pricing

The table below shows the estimated costs using the default configuration. You can customize the configuration later when deploying this solution.

Bitnami WordPress Usage Fee	USD 0.00/mo
Bitnami does not charge a usage fee.	
Infrastructure fee	
VM Instance: 1 shared vCPU + 1.7 GB memory (g1-small)	USD 18.76/mo
Standard Persistent Disk: 10GB	USD 0.47/mo
Sustained use discount	- USD 5.63/mo
Estimated monthly total	USD 13.60/mo

We're currently using USD to calculate costs, which can be changed in the billing setup. Final prices in your bill will be set in accordance with your billing setup, and might be subject to exchange rates.

Price estimates based on 30-day, 24hrs per day usage of the listed resources in the Central US region. The Estimated Monthly Infrastructure Fee calculation may not reflect all Google Cloud Platform IaaS resources actually created or consumed by this product (or the fees charged for such consumption). Bitnami may be able to provide a more accurate estimate of monthly GCP IaaS consumption.

Google Cloud Platform Free Trial

New Google Cloud customers may be eligible for free trial.

[Learn more about Google Cloud pricing & free trial](#)

FIGURE 16.8 Tutorial and support information

Tutorials and documentation

[Access using SSH](#) Configure SSH keys to access the application as the user "bitnami".

[Using SFTP](#) Use this guide to upload files using SFTP.

[MariaDB access credentials](#) Use username "root" and the temporary password to access MariaDB.

[Change your MariaDB root password](#) Change your temporary mariadb root password by following these instructions

[Accessing phpMyAdmin](#) Access phpMyAdmin via an SSH tunnel using this guide.

[Adding plugins with privileges](#) Some plugins need privileged access to install. Edit privileges with this guide.

[Installation directory structure](#) Learn how application files, libraries and configuration files are organized.

Support

Bitnami provides technical support for installation and setup issues through [our support center](#).

[Learn more](#)

Terms of Service

By using this product you agree to the [GCP Marketplace Terms of Service](#) and the terms and conditions of the following software license(s): [End User License Agreement](#).

Além dos parâmetros descritos anteriormente, a página de lançamento também exibirá links para documentação relacionada, como mostrado na Figura 16.12.

Clique no botão Implantar para iniciar a implantação. Isso abrirá o Gerenciador de Implantação e mostrará o progresso da implantação (ver Figura 16.13).

Quando o processo de lançamento for concluído, você verá um resumo sobre a implantação e um botão para lançar o painel administrativo, como mostrado na Figura 16.14.

FIGURE 16.9 Launch a Cloud Marketplace solution from the overview page of the product.

The screenshot shows the product page for 'WordPress Multisite Certified by Bitnami and Automatic'. At the top, there's a Bitnami logo and the product name. Below it, a sub-header reads 'Up-to-date, secure, and ready to run.' Two buttons are present: 'LAUNCH' (in blue) and 'VIEW PAST DEPLOYMENTS'. A navigation bar at the bottom includes 'OVERVIEW' (which is underlined), 'PRICING', and 'SUPPORT'. The main content area starts with a section titled 'Overview' which contains several paragraphs of descriptive text about the WordPress Multisite image. To the right of the overview text is a 'Additional details' sidebar containing various technical specifications and links. At the bottom of the sidebar is a link to 'Add to Service Catalog: Deployment.zip file'.

FIGURE 16.10 The launch page for a WordPress solution in Cloud Marketplace

The screenshot shows the 'New WordPress Certified by Bitnami and Automatic deployment' page. It features a form for setting up a new deployment. The 'Deployment name' field is set to 'wordpress-1'. The 'Zone' dropdown is set to 'us-south1-c'. Under 'Machine type', the 'Machine family' is 'GENERAL-PURPOSE'. The 'Series' dropdown shows 'N2'. Below this, it says 'Powered by Intel Cascade Lake and Ice Lake CPU platforms'. The 'Machine type' dropdown is set to 'n2-standard-2 (2 vCPU, 8 GB memory)'. On the right side, there's a summary table for the 'WordPress Certified by Bitnami and Automatic overview'. The table includes columns for 'Bitnami WordPress Usage Fee' (USD 0.00/mo), 'Infrastructure fee' (USD 83.66/mo), and 'Estimated monthly total' (USD 59.03/mo). A note at the bottom states: 'Price estimates based on 30-day, 24hrs per day usage of the listed resources in the selected region. The Estimated Monthly Infrastructure Fee calculation may not reflect all Google Cloud Platform IaaS resources actually created or consumed by this product (or the fees charged for such consumption). Bitnami may be able to provide a more accurate estimate of monthly GCP IaaS consumption.'

FIGURE 16.11 Additional network parameters

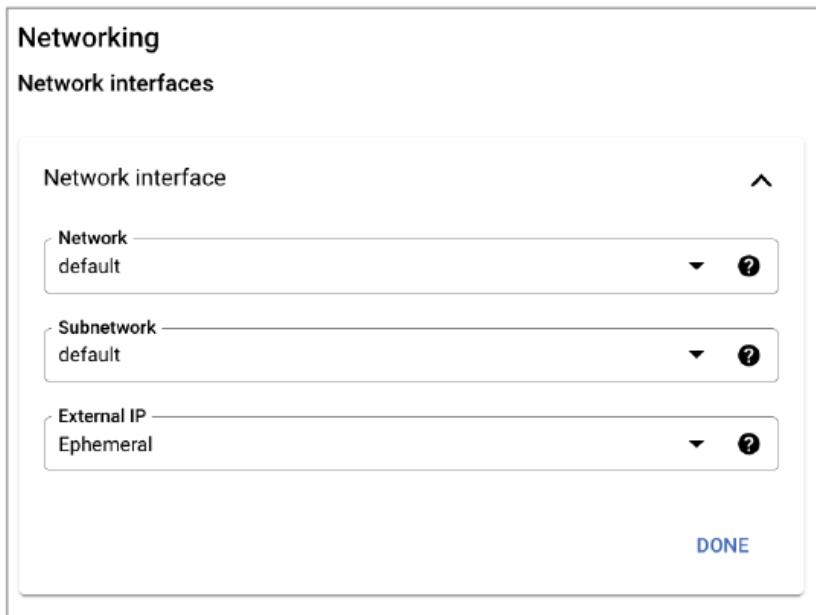


FIGURE 16.12 Links to related documentation are available on the deployment page.

Documentation

- [Access using SSH](#)
- Configure SSH keys to access the application as the user "bitnami".
- [Using SFTP](#)
- Use this guide to upload files using SFTP.
- [MariaDB access credentials](#)
- Use username "root" and the temporary password to access MariaDB.
- [Change your MariaDB root password](#)
- Change your temporary mariadb root password by following these instructions
- [Accessing phpMyAdmin](#)
- Access phpMyAdmin via an SSH tunnel using this guide.
- [Adding plugins with privileges](#)
- Some plugins need privileged access to install. Edit privileges with this guide.
- [Installation directory structure](#)
- Learn how application files, libraries and configuration files are organized.

Terms of Service

By deploying the software or accessing the service you are agreeing to comply with the [Bitnami terms of service](#) , [GCP Marketplace terms of service](#) and the terms of applicable open source software licenses bundled with the software or service. Please review these terms and licenses carefully for details about any obligations you may have related to the software or service. To the limited extent an open source software license related to the software or service expressly supersedes the GCP Marketplace Terms of Service, that open source software license governs your use of that software or service.

By using this product, you understand that certain account and usage information may be shared with Bitnami for the purposes of financial accounting, sales attribution, performance analysis, and support.

Google is providing this software or service "as-is" and any support for this software or service will be provided by Bitnami under their terms of service.

FIGURE 16.13 Cloud Deployment Manager launching WordPress

FIGURE 16.14 Information about the deployed WordPress instance

Construindo Infraestrutura Usando o Cloud Foundation Toolkit

Além de lançar as soluções listadas no Cloud Marketplace, você pode criar seus próprios arquivos de configuração de soluções para que os usuários possam lançar soluções pré-configuradas usando arquivos de configuração do Gerenciador de

Implantação, bem como especificações baseadas em Terraform usando o Cloud Foundation Toolkit. Terraform é uma ferramenta de código aberto para especificar infraestrutura como código. Uma terceira opção, o Conector de Configuração, está disponível para aqueles que desejam gerenciar recursos do Google Cloud usando Kubernetes.

Arquivos de Configuração do Gerenciador de Implantação

Os arquivos de configuração do Gerenciador de Implantação são escritos na sintaxe YAML. Os arquivos de configuração começam com a palavra resources, seguidos por entidades de recurso, que são definidas usando três campos:

- name, que é o nome do recurso.
- type, que é o tipo do recurso, como compute.v1.instance.
- properties, que são pares chave-valor que especificam parâmetros de configuração para o recurso. Por exemplo, uma VM tem propriedades para especificar tipo de máquina, discos e interfaces de rede.

Para informações sobre a sintaxe YAML, veja a documentação oficial em yaml.org. Um exemplo simples definindo uma máquina virtual chamada ace-exam-deployment-vm começa com o seguinte:

```
resources:  
  type: compute.v1.instance  
  name: ace-exam-deployment-vm
```

Em seguida, você pode adicionar propriedades, como o tipo de máquina, configuração de disco e interfaces de rede. A seção de propriedades do arquivo de configuração começa com a palavra properties. Para cada propriedade, há um único par chave-valor ou uma lista de pares chave-valor. A propriedade do tipo de máquina tem um único par chave-valor, com a chave sendo machineType. Discos têm várias propriedades, então seguindo a palavra discs, há uma lista de pares chave-valor. Continuando o exemplo de ace-exam-deployment-vm, a estrutura é a seguinte:

```
resources:  
  type: compute.v1.instance  
  name: ace-exam-deployment-vm  
  properties:  
    machineType: [URL_TIPO_MÁQUINA]
```

Neste exemplo, machineType seria uma URL para uma especificação de recurso da API do Google, como a seguir:

[www.googleapis.com/compute/v1/projects/\[ID_PROJETO\]/zones/us-central1-f/machineTypes/f1-micro](http://www.googleapis.com/compute/v1/projects/[ID_PROJETO]/zones/us-central1-f/machineTypes/f1-micro)

Note que há uma referência a [ID_PROJETO], que você substituiria por um ID de projeto real em um arquivo de configuração. Discos têm propriedades como deviceName e tipo, e Booleanos indicando se o disco é um disco de inicialização ou se deve ser excluído automaticamente. Vamos continuar o exemplo anterior adicionando a especificação do tipo de máquina e algumas propriedades de disco:

resources:

type: compute.v1.instance

name: ace-exam-deployment-vm

properties:

machineType: www.googleapis.com/compute/v1/projects/[ID_PROJETO]/zones/us-central1-f/machineTypes/f1-micro

disks:

- deviceName: boot

type: PERSISTENT

boot: true

autoDelete: true

A Listagem 16.1 mostra o arquivo de configuração completo da documentação do Google Deployment Manager. O seguinte código está disponível em <https://cloud.google.com/deployment-manager/docs/quickstart> (fonte: https://github.com/GoogleCloudPlatform/deploymentmanager-samples/blob/master/examples/v2/quick_start/vm.yaml).

Listing 16.1: examples/v2/quick_start/vm.yaml

```
# Copyright 2016 Google Inc. All rights reserved.
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
# www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#
# Put all your resources under 'resources:'. For each resource, you need:
# -The
# type of resource. In this example, the type is a Compute VM instance.
# -An
internal name for the resource.
# -The
properties for the resource. In this example, for VM instances, you add
# the machine type, a boot disk, network information, and so on.
```

```

#
# For a list of supported resources,
# see https://cloud.google.com/deployment-manager/
docs/configuration/
supported-resource-
types
resources:
-type:
compute.v1.instance
name: quickstart-deployment-
vm
properties:
# The properties of the resource depend on the type of resource. For a
list of properties, see the API reference for the resource.
zone: us-central1-
f
# Replace [MY_PROJECT] with your project ID
machineType: www.googleapis.com/compute/v1/projects/[MY_PROJECT]/
zones/us-central1-
f/
machineTypes/f1-micro
disks:
-deviceName:
boot
type: PERSISTENT
boot: true
autoDelete: true
initializeParams:
# Replace [FAMILY_NAME] with the image family name.
# See a full list of image families at
# https://cloud.google.com/compute/docs/images#os-compute-
support
sourceImage: www.googleapis.com/compute/v1/projects/debian-cloud/
global/images/family/[FAMILY_NAME]
# Replace [MY_PROJECT] with your project ID

networkInterfaces:
-network:
www.googleapis.com/compute/v1/projects/[MY_PROJECT]/
global/networks/default
# Access Config required to give the instance a public IP address
accessConfigs:
-name:
External NAT
type: ONE_TO_ONE_NAT

```

Esta configuração especifica uma implantação chamada quickstart-deployment-vm, que será executada na zona us-central1-f. A implantação usará uma máquina virtual f1-micro executando uma distribuição Debian de Linux. Um endereço IP externo será atribuído. Antes de executar este template, você precisaria substituir [MEU_PROJETO] pelo seu ID de projeto e [NOME_FAMÍLIA] pelo nome de uma família de imagens Debian, como debian-9. Você pode encontrar uma lista de imagens na seção Compute

Engine do Cloud Console na aba Imagens. Você também pode listar imagens usando o comando gcloud compute images list.

Arquivos de Template do Gerenciador de Implantação

Se suas configurações de implantação estão se tornando complicadas, você pode usar templates de implantação. Templates são outro arquivo de texto que você usa para definir recursos e importar esses recursos para arquivos de configuração. Isso permite reutilizar definições de recursos em vários lugares. Os templates podem ser escritos em Python ou Jinja2, uma linguagem de template.

Para informações sobre a sintaxe Jinja2, veja a documentação oficial em <http://jinja.pocoo.org/docs/2.10>.

Como Engenheiro de Nuvem Associado, você deve saber que o Google recomenda usar Python para criar arquivos de template, a menos que os templates sejam relativamente simples, caso em que é apropriado usar Jinja2.

Lançando um Template do Gerenciador de Implantação

Você pode lançar um template de implantação usando o comando gcloud deployment-manager deployments create. Por exemplo, para implantar o template da documentação do Google, use o seguinte:

```
gcloud deployment-manager deployments create quickstart-deployment --config=vm.yaml
```

Você também pode descrever o estado de uma implantação com o comando describe, da seguinte forma:

```
gcloud deployment-manager deployments describe quickstart-deployment
```

Fornecendo um Serviço Implantável

Em grandes empresas, diferentes grupos frequentemente desejam usar o mesmo serviço, como uma aplicação de ciência de dados, para entender os padrões de compra dos clientes. Gerentes de produto em toda a organização podem querer usar isso. Desenvolvedores de software podem criar uma única instância dos recursos da aplicação e ter múltiplos usuários trabalhando com essa única instância. Isso é uma estrutura co-hospedada, que tem algumas vantagens se você tiver uma única equipe de DevOps dando suporte a todos os usuários.

Alternativamente, você poderia permitir que cada usuário ou pequeno grupo de usuários tenha sua própria instância da aplicação. Esta abordagem tem várias vantagens. Usuários poderiam rodar a aplicação em seus próprios projetos, simplificando a alocação de cobranças por recursos, uma vez que o projeto estaria vinculado às contas de cobrança dos usuários. Além disso, os usuários poderiam escalar os recursos para cima ou para baixo conforme necessário para o seu caso de uso.

Uma desvantagem potencial é que os usuários podem não se sentir confortáveis configurando recursos do Google Cloud. O Deployment Manager aborda esse problema tornando relativamente simples implantar uma aplicação e recursos em um processo

repetível. Alguém que possa executar um comando gcloud deployment-manager poderia implantar recursos da aplicação de maneira similar à forma como os usuários implantam aplicações a partir do Cloud Marketplace.

Cloud Foundation Toolkit

O Cloud Foundation Toolkit é um projeto de código aberto que fornece templates de infraestrutura como código usando templates do Deployment Manager e Terraform.

O Cloud Foundation Toolkit inclui blueprints, que são pacotes de especificação de configuração implantável, bem como políticas para implementar uma solução para uma determinada classe de problemas. Esses blueprints encapsulam as melhores práticas para configurar infraestrutura e conceder acesso a recursos. Blueprints estão disponíveis para Terraform e Kubernetes. Os blueprints de Kubernetes são usados com o Config Connector. Para exemplos de blueprints, veja <https://cloud.google.com/docs/terraform/blueprints/terraform-blueprints>.

Além dos blueprints projetados para resolver necessidades amplas, como a implantação de um data warehouse, templates também estão disponíveis para configurar recursos específicos do serviço Google Cloud. Por exemplo, a Listagem 16.2 mostra um template para criar uma máquina virtual.

```
Listing 16.2: https://github.com/terraform-google-modules/terraform-google-vm/blob/master/modules/compute_instance/main.tf
/**
 * Copyright 2018 Google LLC
 * Licensed under the Apache License, Version 2.0 (the "License");
 * you may not use this file except in compliance with the License.
 * You may obtain a copy of the License at
 *
 * www.apache.org/licenses/LICENSE-2.0
 *
 * Unless required by applicable law or agreed to in writing, software
 * distributed under the License is distributed on an "AS IS" BASIS,
 * WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 * See the License for the specific language governing permissions and
 * limitations under the License.
 */
locals {
  hostname = var.hostname == "" ? "default" : var.hostname
  num_instances = length(var.static_ips) == 0 ? var.num_instances :
    length(var.static_ips)
  # local.static_ips is the same as var.static_ips with a dummy element
  # appended
  # at the end of the list to work around "list does not have any elements
  # so cannot
  # determine type" error when var.static_ips is empty
  static_ips = concat(var.static_ips, ["NOT_AN_IP"])
  project_id = length(regexall("/projects/([^\/*]", var.instance_template)) >
    0 ? flatten(regexall("/projects/([^\/*]", var.instance_template))[0] : null
```

```

# When no network or subnetwork has been defined, we want to use the
# settings from
# the template instead.
network_interface = length(format("%s%s", var.network, var.subnetwork)) == 0
? [] : [1]
#####
# Data Sources
#####
data "google_compute_zones" "available" {
project = local.project_id
region = var.region}
#####
# Instances #####
resource "google_compute_instance_from_template" "compute_instance" {
provider = google
count = local.num_instances
name = var.add_hostname_suffix ? format("%s%s%s", local
.hostname, var.hostname_suffix_separator, format("%03d", count.index + 1)) :
local.hostname
project = local.project_id
zone = var.zone == null ? data.google_compute_zones
.available.names[count.index % length(data.google_compute_zones.available
.names)] : var.zone
deletion_protection = var.deletion_protection
dynamic "network_interface" {
for_each = local.network_interface
content {
network = var.network
subnetwork = var.subnetwork
subnetwork_project = var.subnetwork_project
network_ip = length(var.static_ips) == 0 ? "" : element(local
.static_ips, count.index)
dynamic "access_config" {
for_each = var.access_config
content {
nat_ip = access_config.value.nat_ip
network_tier = access_config.value.network_tier}}
dynamic "alias_ip_range" {
for_each = var.alias_ip_ranges
content {
ip_cidr_range = alias_ip_range.value.ip_cidr_range
subnetwork_range_name = alias_ip_range.value.subnetwork_range_name}}}}
source_instance_template = var.instance_template}

```

Config Connector

O Config Connector é um add-on do Kubernetes que permite gerenciar recursos do Google Cloud por meio do Kubernetes. Isso é útil para aqueles que já gerenciaram recursos do Kubernetes usando configurações do Kubernetes e desejam estender o escopo dessas ferramentas para incluir recursos do Google Cloud. O Config Connector fornece

uma coleção de definições de recursos personalizados do Kubernetes (CRDs) e controladores para gerenciar recursos do Google Cloud.

Para instalar o Config Connector, você passa um parâmetro para o comando gcloud container clusters create especificando o add-on ConfigConnector. Por exemplo:

```
gcloud container clusters create ace-gke-cluster1 \
--addons ConfigConnector \
--workload-pool=ace-project-dw1 \
--logging=SYSTEM \
--monitoring=SYSTEM
```

Para usar o Config Connector, você terá que habilitar a Identidade de Carga de Trabalho, uma maneira de vincular identidades do IAM a identidades do Kubernetes. Você também precisará habilitar o monitoramento do Kubernetes Engine e usar uma versão suportada do Kubernetes. Configurações do Config Connector são aplicadas usando kubectl.

Para mais sobre soluções Config Connector, veja <https://github.com/GoogleCloudPlatform/cloud-foundation-toolkit/tree/master/config-connector/solutions>.

Resumo

Cloud Marketplace e Cloud Deployment Manager são projetados para facilitar a implantação de recursos no Google Cloud. Cloud Marketplace é onde fornecedores terceirizados podem oferecer aplicações implantáveis baseadas em software proprietário ou de código aberto. Quando uma aplicação é implantada a partir do Cloud Marketplace, recursos como VMs, buckets de armazenamento e discos persistentes são criados automaticamente sem intervenção humana adicional. O Deployment Manager dá aos engenheiros de nuvem a capacidade de definir arquivos de configuração que descrevem os recursos que gostariam de implantar. Engenheiros de nuvem podem então usar comandos gcloud para implantar os recursos e listar seu status. O Deployment Manager é especialmente útil em organizações onde você deseja implantar recursos facilmente sem exigir que os usuários desses recursos entendam os detalhes de como configurar recursos do Google Cloud.

O Cloud Foundation Toolkit fornece templates e blueprints que codificam as melhores práticas para implantar soluções e recursos individuais no Google Cloud. O add-on Config Connector para Kubernetes permite gerenciar recursos do Google Cloud usando o Kubernetes.

Essenciais do Exame

Entenda como buscar soluções usando a seção Cloud Marketplace do Cloud Console. Você pode usar filtros para restringir sua pesquisa a tipos específicos de soluções, como sistemas operacionais e ferramentas de desenvolvedor. Pode haver

múltiplas opções para uma única aplicação, como o WordPress. Isso ocorre porque vários fornecedores oferecem configurações. Revise a descrição de cada uma para entender qual se ajusta melhor às suas necessidades.

Saiba como implantar uma solução no Cloud Marketplace. Entenda como configurar uma implantação do Cloud Marketplace no Cloud Console. Entenda que, quando você lança uma solução, pode ser solicitado para configurações específicas da aplicação. Por exemplo, com o WordPress, pode ser solicitado para instalar o phpMyAdmin. Você também pode ter a oportunidade de configurar atributos de configuração comuns, como o tipo de máquina e o tipo de disco de inicialização.

Entenda como usar a seção do Deployment Manager do console para monitorar a implantação. Pode levar alguns minutos desde o momento em que você lança uma configuração até o momento em que ela está pronta para uso. Note que, uma vez que a aplicação está pronta, você pode ser solicitado por informações adicionais, como um nome de usuário e senha para fazer login.

Saiba que o Deployment Manager é um serviço do Google Cloud para criar arquivos de configuração que definem recursos para usar com uma aplicação. Estes arquivos de configuração usam sintaxe YAML. Eles são compostos por especificações de recursos que usam pares chave-valor para definir propriedades do recurso.

Saiba que recursos em um arquivo de configuração são definidos usando um nome, tipo e conjunto de propriedades. As propriedades variam de acordo com o tipo. O tipo de máquina pode ser definido usando apenas uma URL que aponta para um tipo de máquina disponível em uma região. Discos têm várias propriedades, incluindo um nome de dispositivo, um tipo e se o disco é um disco de inicialização.

Saiba que você pode usar templates com arquivos de configuração. Se seus arquivos de configuração estão ficando longos ou complicados, você pode modularizá-los usando templates. Templates definem recursos e podem ser importados em outros templates. Templates são arquivos de texto escritos em Jinja2 ou Python.

Saiba como lançar um arquivo de configuração de implantação usando o comando `gcloud deployment-manager deployments create`. Você pode revisar o status de uma implantação usando o comando `gcloud deployment-manager deployments describe`.

Conheça o propósito do Cloud Foundation Toolkit e do Config Connector. Cloud Foundation Toolkit é um projeto de código aberto com blueprints e configurações de exemplo que capturam as melhores práticas recomendadas pelo Google Cloud para implantar soluções. Config Connector é um add-on do Kubernetes para gerenciar recursos do Google Cloud a partir do Kubernetes.

Questões de Revisão

Você pode encontrar as respostas no Apêndice.

1. Quais são as categorias de soluções do Cloud Marketplace?
 - A. Apenas conjuntos de dados
 - B. Apenas sistemas operacionais
 - C. Apenas ferramentas de desenvolvedor e sistemas operacionais
 - D. Conjuntos de dados, sistemas operacionais e ferramentas de desenvolvedor
2. Você quer usar o Terraform para gerenciar infraestrutura como código e também gostaria de seguir as melhores práticas recomendadas pelo Google Cloud. O que você usaria para começar a implementar tal solução?
 - A. Cloud Deployment Manager
 - B. Cloud Foundation Toolkit
 - C. Config Connector
 - D. Cloud Build
3. Onde você navega para lançar uma solução do Cloud Marketplace?
 - A. Página de visão geral da solução
 - B. Página principal do Cloud Marketplace
 - C. Serviços de Rede
 - D. Nenhuma das opções acima
4. Você quer identificar rapidamente o conjunto de sistemas operacionais disponíveis no Cloud Marketplace. Qual destes passos ajudaria com isso?
 - A. Usar o Google Search para buscar na web uma listagem.
 - B. Usar filtros no Cloud Marketplace.
 - C. Percorrer a lista de soluções exibidas na página inicial do Cloud Marketplace.
 - D. Não é possível filtrar para sistemas operacionais.
5. Você quer usar o Cloud Marketplace para implantar um site WordPress. Você nota que há mais de uma opção WordPress. Por que isso?
 - A. É um erro. Envie um chamado para o suporte do Google.
 - B. Vários fornecedores podem oferecer a mesma aplicação.
 - C. É um erro. Envie um chamado para os fornecedores.
 - D. Você nunca verá tal opção.

6. Você usou o Cloud Marketplace para implantar um site WordPress e agora gostaria de implantar um banco de dados. Você percebe que a página de configuração dos bancos de dados é diferente da usada com o WordPress. Por que isso?

 - A. É um erro. Envie um chamado para o suporte do Google.
 - B. Você navegou para um subformulário diferente do Cloud Marketplace.
 - C. As propriedades de configuração são baseadas na aplicação que você está implantando e serão diferentes dependendo de qual aplicação você está implantando.
 - D. Isso não pode acontecer.
7. Seu gerente pediu para você implantar um site WordPress. Você espera um tráfego intenso, e seu gerente quer garantir que a VM hospedando o site WordPress tenha recursos suficientes. Quais recursos você pode configurar ao lançar um site WordPress usando o Cloud Marketplace?

 - A. Tipo de máquina
 - B. Tipo de disco
 - C. Tamanho do disco
 - D. Todos os itens acima
8. Você gostaria de definir como código a configuração de um conjunto de recursos de aplicativos. Qual é o serviço do Google Cloud para criar recursos usando um arquivo de configuração composto por especificações de recursos definidas em sintaxe YAML?

 - A. Compute Engine
 - B. Deployment Manager
 - C. Marketplace Manager
 - D. Marketplace Deployer
9. Qual formato de arquivo é usado para definir arquivos de configuração do Deployment Manager?

 - A. XML
 - B. JSON
 - C. CSV
 - D. YAML
10. Um arquivo de configuração do Deployment Manager começa com qual palavra?

 - A. deploy
 - B. resources (recursos)
 - C. properties (propriedades)

D. YAML

11. Quais dos seguintes são usados para definir um recurso em um arquivo de configuração do Cloud Deployment Manager?
 - A. Apenas Tipo
 - B. Apenas Propriedades
 - C. Apenas Nome e Tipo
 - D. Tipo, Propriedades e Nome
12. Quais propriedades podem ser definidas ao definir um disco em uma VM?
 - A. Apenas um nome de dispositivo
 - B. Um Booleano indicando um disco de inicialização e um Booleano indicando autodeleção
 - C. Apenas um Booleano indicando autodeleção
 - D. Um nome de dispositivo, um Booleano indicando um disco de inicialização e um Booleano indicando autodeleção
13. Você precisa verificar quais imagens estão disponíveis na zona na qual deseja implantar uma VM. Qual comando você usaria?
 - A. gcloud compute images list
 - B. gcloud images list
 - C. gsutil compute images list
 - D. gcloud compute list images
14. Você quer usar um arquivo de template com o Deployment Manager. Você espera que o arquivo seja complicado. Qual linguagem você usaria?
 - A. Jinja2
 - B. Ruby
 - C. Go
 - D. Python
15. Qual comando lança uma implantação?
 - A. gcloud deployment-manager deployments create
 - B. gcloud cloud-launcher deployments create
 - C. gcloud deployment-manager deployments launch
 - D. gcloud cloud-launcher deployments launch

16. Um engenheiro de DevOps está notando um pico na utilização de CPU em seus servidores. Você explica que acabou de lançar uma implantação. Você gostaria de mostrar ao engenheiro de DevOps os detalhes de uma implantação que acabou de lançar. Qual comando você usaria?
- A. gcloud cloud-launcher deployments describe
 - B. gcloud deployment-manage deployments list
 - C. gcloud deployment-manager deployments describe
 - D. gcloud cloud-launcher deployments list
17. Se você expandir o link Mais na seção de Rede ao implantar uma solução do Cloud Marketplace, o que você poderá configurar?
- A. Endereços IP
 - B. Cobrança
 - C. Controles de acesso
 - D. Tipo de máquina personalizado
18. Quais são os tipos de licença referenciados no Cloud Marketplace?
- A. Apenas gratuito
 - B. Gratuito e tarifa fixa por hora apenas
 - C. Gratuito e traga sua própria licença (BYOL) apenas
 - D. Gratuito, tarifa fixa por hora, taxas de uso e traga sua própria licença (BYOL)
19. Qual tipo de licença adicionará cobranças à sua fatura do Google Cloud quando usar o Cloud Marketplace com este tipo de licença?
- A. Gratuito
 - B. Tarifa fixa por hora e taxas de uso
 - C. BYOL
 - D. Chargeback
20. Você está implantando uma aplicação do Cloud Marketplace que inclui um stack LAMP. Que software isso implantará?
- A. Servidor Apache e Linux apenas
 - B. Apenas Linux
 - C. MySQL e Apache apenas
 - D. Apache, MySQL, Linux e PHP

Capítulo 17

Configurando Acesso e Segurança

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 5.1 Gerenciando Gerenciamento de Identidade e Acesso (IAM)
- ✓✓ 5.2 Gerenciando contas de serviço
- ✓✓ 5.3 Visualizando logs de auditoria

Engenheiros do Google Cloud podem esperar gastar uma quantidade significativa de tempo trabalhando com controles de acesso. Este capítulo fornece instruções sobre como realizar várias tarefas comuns, incluindo gerenciar atribuições de gerenciamento de identidade e acesso (IAM), criar funções personalizadas, gerenciar contas de serviço e visualizar logs de auditoria.

É importante saber que a maneira preferida de atribuir permissões a usuários, grupos e contas de serviço é através do sistema IAM. No entanto, o Google Cloud nem sempre teve o IAM. Antes disso, as permissões eram concedidas usando o que agora são conhecidas como funções básicas, que são bastante genéricas.

Funções básicas podem ter mais permissões do que você deseja que uma identidade tenha. Você pode restringir permissões usando escopos. Neste capítulo, descreveremos como usar funções básicas e escopos, bem como o IAM. Daqui para frente, é uma melhor prática usar o IAM para controle de acesso.

Gerenciando Gerenciamento de Identidade e Acesso

Quando você trabalha com o IAM, há algumas tarefas comuns que você precisa realizar:

- Visualizando atribuições de IAM da conta
- Atribuindo funções de IAM
- Definindo funções personalizadas

Vamos ver como realizar cada uma dessas tarefas.

Visualizando Atribuições de IAM da Conta

Você pode visualizar atribuições de IAM da conta no Cloud Console, navegando até a seção IAM & Admin. Nessa seção, selecione IAM no menu de navegação para exibir a página mostrada na Figura 17.1. O exemplo na figura mostra uma lista de identidades filtrada por nome do membro.

Neste exemplo, o usuário dan@sullivanlearninggroup.com tem três funções: Administrador de Recursos da Organização do Compute, Administrador da Organização e Proprietário. Admin do App Engine e Admin do BigQuery são funções de IAM predefinidas. Proprietário é uma função básica.

FIGURE 17.1 Permissions listing filtered by member

Type	Principal	Name	Role
<input type="checkbox"/>	388947348090-compute@developer.gserviceaccount.com	Compute Engine default service account	Cloud Data Fusion Runner
<input type="checkbox"/>			Editor
<input type="checkbox"/>	388947348090@cloudbuild.gserviceaccount.com		Cloud Build Service Account
<input type="checkbox"/>	388947348090@cloudservices.gserviceaccount.com	Google APIs Service Agent	Editor
<input type="checkbox"/>	dan@sullivanlearninggroup.com	Dan Sullivan	Compute Organization Resource Admin
			Organization Administrator
			Owner
<input type="checkbox"/>	scenic-energy-335022@appspot.gserviceaccount.com	App Engine default service account	Owner

As funções básicas eram usadas antes do IAM. Existem três funções básicas: Proprietário, Editor e Visualizador. Visualizadores têm permissão para realizar operações somente leitura. Editores têm permissões de visualizador e permissão para modificar uma entidade. Proprietários têm permissões de editor e podem gerenciar funções e permissões em uma entidade. Proprietários também podem configurar a cobrança para um projeto.

As funções do IAM são coleções de permissões. Elas são personalizadas para fornecer às identidades apenas as permissões de que precisam para realizar uma tarefa, e nada mais. Para ver uma lista de usuários atribuídos a uma função (veja a Figura 17.2), clique na aba Funções na página do IAM.

FIGURE 17.2 List of identities assigned to Cloud Build Service Account and Cloud Data Fusion Runner roles

Role / Principal	Name	Inheritance
<input type="checkbox"/>	Cloud Build Service Account (1)	
<input type="checkbox"/>	388947348090@cloudbuild.gserviceaccount.com	
<input type="checkbox"/>	Cloud Data Fusion Runner (1)	
<input type="checkbox"/>	388947348090-compute@developer.gserviceaccount.com	Compute Engine default service account

Esta página mostra uma lista de funções com o número de identidades atribuídas a essa função entre parênteses. Clique na seta ao lado do nome de uma função para exibir uma lista de identidades com essa função.

Você também pode ver uma lista de usuários e funções atribuídas em um projeto usando o comando `gcloud projects get-iam-policy`. Por exemplo, para listar funções atribuídas a usuários em um projeto com o ID do projeto `ace-exam-project`, use isso:

```
gcloud projects get-iam-policy ace-exam-project
```

As funções predefinidas são agrupadas por serviço. Por exemplo, o App Engine tem cinco funções:

■■ Administrador do App Engine, que concede permissão de leitura, escrita e modificação para configurações de aplicativos e. O nome da função usado nos comandos `gcloud` é `roles/appengine.appAdmin`.

■■ Administrador de Serviço do App Engine, que concede acesso somente leitura às configurações e acesso de escrita às configurações de nível de módulo e de versão. O nome da função usado nos comandos `gcloud` é `roles/appengine.serviceAdmin`.

■■ Implementador do App Engine, que concede acesso somente leitura à configuração do aplicativo e às configurações e acesso de escrita para criar novas versões. Usuários apenas com a função de Implementador do App Engine não podem modificar ou deletar versões existentes. O nome da função usado nos comandos `gcloud` é `roles/appengine.deployer`.

■■ Visualizador do App Engine, que concede acesso somente leitura à configuração do aplicativo e às configurações. O nome da função usado nos comandos `gcloud` é `roles/appengine.appViewer`.

■■ Visualizador de Código do App Engine, que concede acesso somente leitura a todas as configurações do aplicativo, configurações e código fonte implantado. O nome da função usado nos comandos `gcloud` é `roles/appengine.codeViewer`.

Embora você não precise conhecer todas elas, é útil revisar as funções predefinidas para entender os padrões de como elas são definidas. Para mais detalhes, veja a documentação do Google Cloud em <https://cloud.google.com/iam/docs/understanding-roles>.

Atribuindo Funções do IAM a Contas e Grupos

Para adicionar funções do IAM a contas e grupos, navegue até a seção IAM & Admin do console. Selecione IAM no menu. Clique no link Adicionar no topo para exibir uma página como a mostrada na Figura 17.3.

Especifique o nome de um usuário ou grupo no campo rotulado Novos Principais. Clique em Selecionar Uma Função para adicionar uma função. Você pode adicionar várias funções. Quando você clica na seta para baixo no campo Selecionar Uma Função, você verá uma lista de serviços e suas funções associadas. Você pode escolher as funções dessa lista. Veja a Figura 17.4 para um exemplo de um subconjunto da lista, mostrando as funções para o BigQuery.

FIGURE 17.3 The Add option in IAM opens this page, where you can assign one or more roles to users or groups.

Add principals to "My First Project"

Add principals, roles to "My First Project" project

Enter one or more principals below. Then select a role for these principals to grant them access to your resources. Multiple roles allowed. [Learn more](#)

New principals *

Select a role *

Condition [Add condition](#)

+ ADD ANOTHER ROLE

SAVE CANCEL

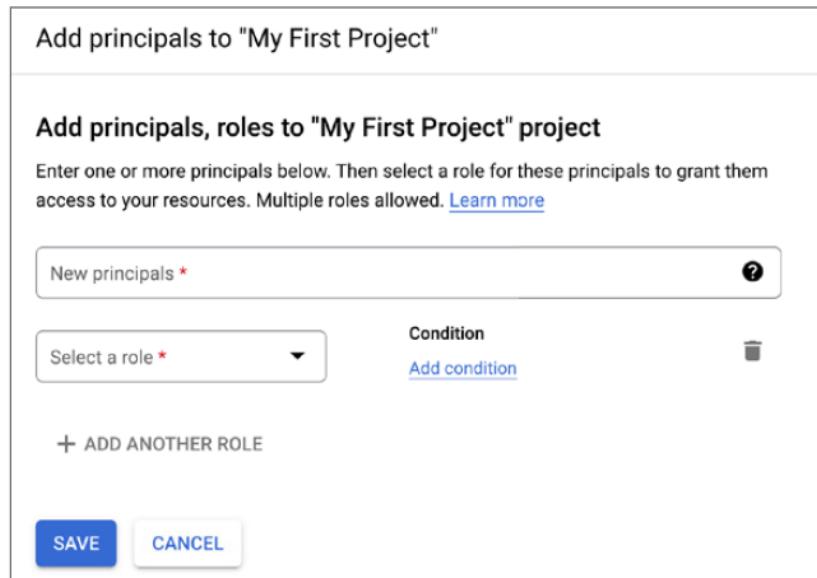
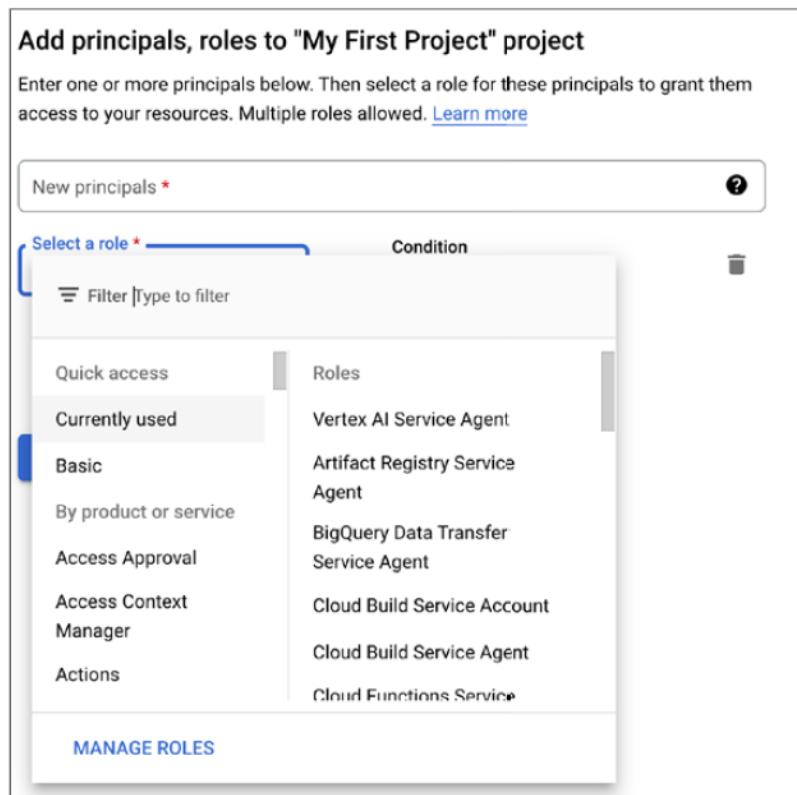


FIGURE 17.4 The drop-down list in the Select A Role field shows available roles grouped by service.



Se você quiser saber quais das permissões detalhadas são concedidas ao atribuir uma função, você pode listar essas permissões na linha de comando ou no console.

Você também pode ver quais permissões são atribuídas a uma função usando o comando gcloud iam roles describe. Por exemplo, a Figura 17.5 mostra a lista de permissões no papel de Administrador de Banco de Dados do Spanner.

FIGURE 17.5 A partial listing of permissions using the gcloud iam roles describe command

```
dansullivanblk@cloudshell:~ (gdg-project-294122)$ gcloud iam roles describe roles/spanner.databaseAdmin
description: Full control of Cloud Spanner databases.
etag: AA==
includedPermissions:
- monitoring.timeSeries.list
- resourcemanager.projects.get
- resourcemanager.projects.list
- spanner.databaseOperations.cancel
- spanner.databaseOperations.delete
- spanner.databaseOperations.get
- spanner.databaseOperations.list
- spanner.databases.beginOrRollbackReadWriteTransaction
- spanner.databases.beginPartitionedDmlTransaction
- spanner.databases.beginReadOnlyTransaction
- spanner.databases.create
- spanner.databases.drop
- spanner.databases.get
- spanner.databases.getDdl
- spanner.databases.getIamPolicy
- spanner.databases.list
- spanner.databases.partitionQuery
- spanner.databases.partitionRead
- spanner.databases.read
- spanner.databases.select
```

Você também pode usar o Cloud Console para visualizar permissões. Navegue até a seção IAM & Admin e selecione Funções no menu. Isso exibirá uma lista de funções. Clique na caixa de seleção ao lado de um nome de função para exibir uma lista de permissões à direita, como mostrado na Figura 17.6 para o Administrador do Cloud SQL.

Você pode atribuir funções a um membro em um projeto usando o seguinte comando:

```
gcloud projects add-iam-policy-binding [NOME-DO-RECURSO] --member=user:[EMAIL-DO-USUÁRIO] --role=[ID-DA-FUNÇÃO]
```

Por exemplo, para conceder a função básica de Editor a um usuário identificado por jane@aceexam.com, você poderia usar isto:

```
gcloud projects add-iam-policy-binding ace-exam-project --member=user:jane@aceexam.com --role='roles/editor'
```

FIGURE 17.6 Using Cloud Console to view a partial listing of permissions available for Cloud SQL Admin

The screenshot shows the Google Cloud Console interface for managing roles. At the top, there's a back arrow, the title "Cloud SQL Admin", and two buttons: "+ EDIT ROLE" and "CREATE FROM ROLE". Below this, there are two rows of information: "ID" followed by "roles/cloudsql.admin" and "Role launch stage" followed by "General Availability". A section titled "Description" contains the text "Full control of Cloud SQL resources.". The most prominent part of the page is a large list titled "71 assigned permissions" which includes the following items:

- cloudsql.backupRuns.create
- cloudsql.backupRuns.delete
- cloudsql.backupRuns.get
- cloudsql.backupRuns.list
- cloudsql.databases.create
- cloudsql.databases.delete
- cloudsql.databases.get
- cloudsql.databases.list
- cloudsql.databases.update
- cloudsql.instances.addServerCa
- cloudsql.instances.clone
- cloudsql.instances.connect
- cloudsql.instances.create
- cloudsql.instances.createTagBinding
- cloudsql.instances.delete
- cloudsql.instances.deleteTagBinding
- cloudsql.instances.demoteMaster
- cloudsql.instances.export
- cloudsql.instances.failover
- cloudsql.instances.get
- cloudsql.instances.import
- cloudsql.instances.list
- cloudsql.instances.listEffectiveTags
- cloudsql.instances.listServerCas
- cloudsql.instances.listTagBindings
- cloudsql.instances.login
- cloudsql.instances.promoteReplica
- cloudsql.instances.resetSslConfig
- cloudsql.instances.restart
- cloudsql.instances.restoreBackup
- cloudsql.instances.rotateServerCa
- cloudsql.instances.startReplica
- cloudsql.instances.stopReplica
- cloudsql.instances.truncateLog
- cloudsql.instances.update

Funções do IAM Suportam o Princípio do Menor Privilégio e Separação de Funções

Duas melhores práticas de segurança são atribuir o menor privilégio e manter uma separação de funções. O princípio do menor privilégio diz que você concede apenas o menor conjunto de permissões necessário para um usuário ou conta de serviço realizar suas tarefas necessárias. Por exemplo, se os usuários podem fazer tudo o que precisam fazer com apenas permissão de leitura em um banco de dados, então eles não devem ter permissão de escrita.

No caso da separação de funções, a ideia é que um único usuário não deve ser capaz de realizar múltiplas operações sensíveis que juntas poderiam apresentar um risco. Em domínios de alto risco, como finanças ou defesa, você não gostaria que um desenvolvedor pudesse modificar uma aplicação e implementar essa mudança em produção sem revisão. Um engenheiro malicioso, por exemplo, poderia modificar o código em uma aplicação financeira para suprimir o registro em log quando fundos são transferidos para uma conta bancária controlada pelo engenheiro malicioso. Se esse engenheiro colocasse esse código em produção, poderia passar algum tempo antes que os auditores descobrissem que o registro em log foi suprimido e que pode ter havido transações fraudulentas.

As funções do IAM suportam o menor privilégio ao atribuir permissões mínimas a funções predefinidas. Também suporta a separação de funções ao permitir que alguns usuários tenham a capacidade de alterar o código e outros de implementar o código.

Outra prática comum de segurança é a defesa em profundidade, que aplica múltiplos controles de segurança sobrepostos. Essa também é uma prática que deve ser adotada. O IAM pode ser aplicado como uma das camadas de defesa.

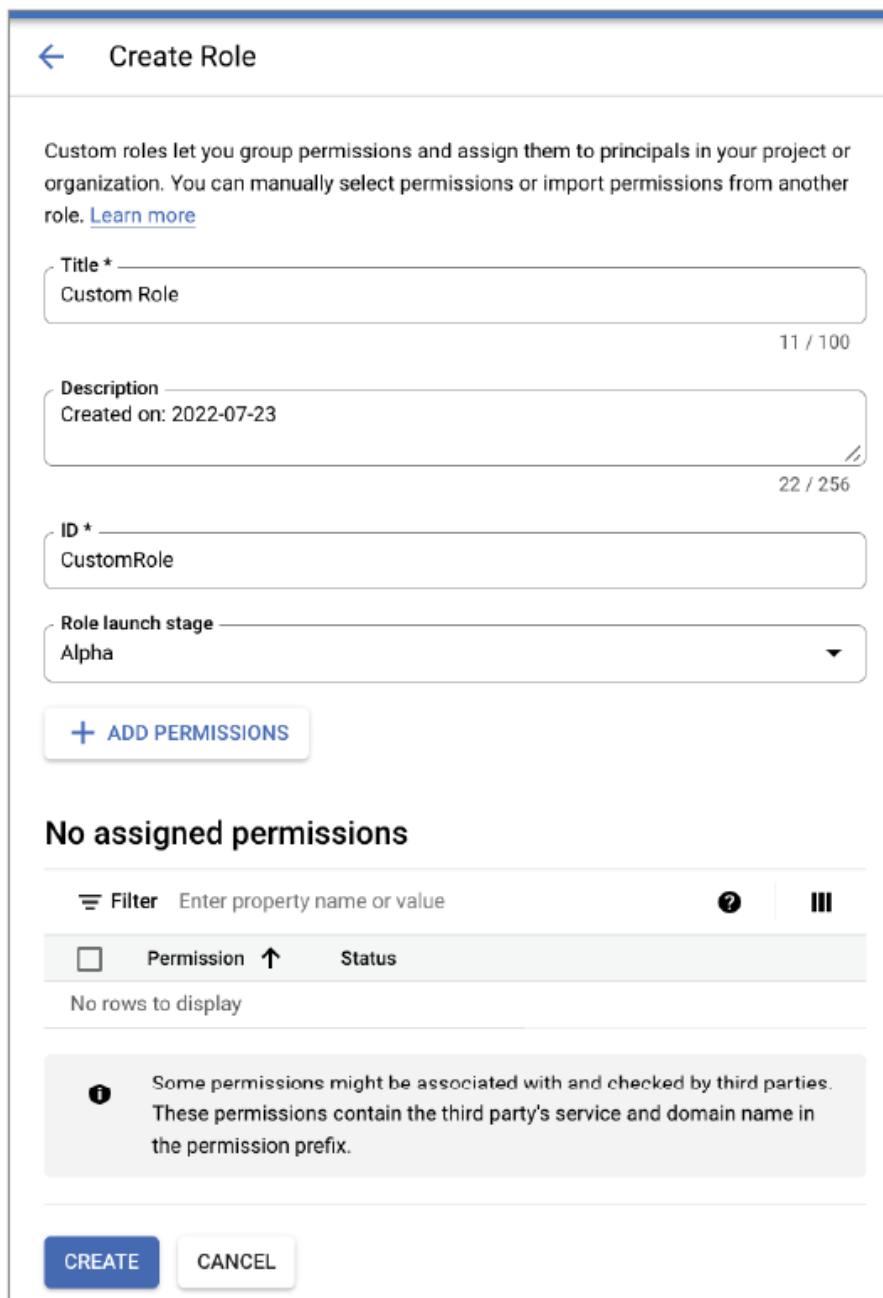
Definindo Funções Customizadas do IAM

Se o conjunto de funções predefinidas do IAM não atender às suas necessidades, você pode definir uma função customizada. Para definir uma função customizada no Cloud Console, navegue até a opção Funções na seção IAM & Admin do console. Clique no link Criar Função no topo da página. Isso exibirá uma página como a mostrada na Figura 17.7.

Nesta página, você pode especificar um nome para a função customizada, uma descrição, um identificador, uma fase de lançamento e um conjunto de permissões. As opções de fase de lançamento são as seguintes: Alpha, Beta, Disponibilidade Geral e Desabilitado.

Você pode clicar em Adicionar Permissões para exibir uma lista de permissões. A lista na Figura 17.8 é filtrada para incluir apenas permissões no papel de Administrador do Cloud SQL.

FIGURE 17.7 Creating a role in Cloud Console



Embora a lista inclua todas as permissões na função, nem todas as permissões estão disponíveis para uso em uma função customizada. Quando uma permissão não está disponível, seu status é listado como Não Suportado. Permissões que estão disponíveis para uso são listadas como Suportadas, então, no exemplo, todas as outras permissões estão disponíveis. Marque as caixas ao lado das permissões que você deseja incluir em sua função customizada. Clique em Adicionar para retornar à página Criar Função, onde a lista de permissões agora incluirá as permissões que você selecionou (veja a Figura 17.9).

FIGURE 17.8 List of available permissions filtered by role

The screenshot shows a 'Cloud SQL Admin' role selected in a dropdown filter. A table lists 10 permissions under the 'cloudsql.' prefix, all marked as 'Supported'. The table has columns for 'Permission' and 'Status'. At the bottom, it says '1 - 10 of 71' with navigation arrows, and at the very bottom are 'CANCEL' and 'ADD' buttons.

<input type="checkbox"/> Permission ↑	Status
cloudsql.backupRuns.create	Supported
cloudsql.backupRuns.delete	Supported
cloudsql.backupRuns.get	Supported
cloudsql.backupRuns.list	Supported
cloudsql.databases.create	Supported
cloudsql.databases.delete	Supported
cloudsql.databases.get	Supported
cloudsql.databases.list	Supported
cloudsql.databases.update	Supported
cloudsql.instances.addServerCa	Supported

Você também pode definir uma função personalizada usando o comando gcloud iam roles create. A estrutura desse comando é a seguinte:

```
gcloud iam roles create [ID-DA-FUNÇÃO] --project [ID-DO-PROJETO] --
title=[TÍTULO-DA-FUNÇÃO] --description=[DESCRIÇÃO-DA-FUNÇÃO] --
permissions=[LISTA-DE-PERMISSÕES] --stage=[FASE-DE-LANÇAMENTO]
```

Por exemplo, para criar uma função que tenha apenas permissão de atualização de aplicativos do App Engine, você poderia usar o seguinte comando:

```
gcloud iam roles create customAppEngine1 --project ace-exam-project --title='Custom
Update      App      Engine'      --description='Custom      update'      --
permissions=appengine.applications.update --stage=alpha
```

FIGURE 17.9 The permissions section of the Create Role page with permissions added

The screenshot shows the 'Create Role' page with the following details:

- Title ***: Custom Role (11 / 100 characters)
- Description**: Created on: 2022-07-23 (22 / 256 characters)
- ID ***: CustomRole
- Role launch stage**: Alpha
- ADD PERMISSIONS** button
- 6 assigned permissions** table:

	Permission ↑	Status
<input checked="" type="checkbox"/>	cloudsql.backupRuns.create	Supported
<input checked="" type="checkbox"/>	cloudsql.backupRuns.delete	Supported
<input checked="" type="checkbox"/>	cloudsql.backupRuns.get	Supported
<input checked="" type="checkbox"/>	cloudsql.backupRuns.list	Supported
<input checked="" type="checkbox"/>	cloudsql.databases.get	Supported
<input checked="" type="checkbox"/>	cloudsql.databases.list	Supported
- INFO** message: Some permissions might be associated with and checked by third parties. These permissions contain the third party's service and domain name in the permission prefix.
- SHOW ADDED AND REMOVED PERMISSIONS** link
- CREATE** and **CANCEL** buttons

Gerenciando Contas de Serviço

Contas de serviço são usadas para fornecer identidades independentes de usuários humanos. Contas de serviço são identidades às quais podem ser atribuídas funções. Contas de serviço são designadas a VMs, que então usam as permissões disponíveis para as contas de serviço para realizar tarefas.

Três coisas que engenheiros de nuvem são esperados saber como fazer são trabalhar com escopos, atribuir contas de serviço a VMs e conceder acesso a uma conta de serviço a outro projeto.

Gerenciando Contas de Serviço com Escopos

Escopos são permissões concedidas a uma VM para realizar alguma operação. Escopos autorizam o acesso a métodos de API. A conta de serviço atribuída a uma VM tem funções associadas a ela. Para configurar controles de acesso para uma VM, você precisará configurar tanto funções de IAM quanto escopos.

Já discutimos como gerenciar funções de IAM, então agora voltaremos nossa atenção para os escopos. Um escopo é especificado usando uma URL que começa com `www.googleapis.com/auth` e é seguido por permissão em um recurso. Por exemplo, o escopo que permite a uma VM inserir dados no BigQuery é o seguinte:

www.googleapis.com/auth/bigquery.insertdata

O escopo que permite visualizar dados no Cloud Storage é o seguinte:

www.googleapis.com/auth/devstorage.read_only

E para escrever em logs do Compute Engine, use isto:

www.googleapis.com/auth/logging.write

Uma instância só pode realizar operações permitidas tanto pelas funções de IAM atribuídas à conta de serviço quanto pelos escopos definidos na instância. Por exemplo, se uma função concede apenas acesso somente leitura ao Cloud Storage, mas um escopo permite acesso de escrita, então a instância não será capaz de escrever no Cloud Storage.

Para definir escopos em uma instância, navegue até a página da instância de VM no Cloud Console. Pare a instância se ela estiver em execução. Na página Detalhe da Instância, clique no link Editar. No meio da página de edição, você verá a seção Escopos de Acesso, como mostrado na Figura 17.10.

As opções são Permitir Acesso Padrão, Permitir Acesso Total a Todas as APIs do Cloud e Definir Acesso Para Cada API. O acesso padrão geralmente é suficiente. Se você não tiver certeza do que definir, pode escolher Permitir Acesso Total, mas certifique-se de atribuir funções de IAM para limitar o que a instância pode fazer. Se você quiser escolher escopos individualmente, escolha Definir Acesso Para Cada API. Isso exibirá uma lista de serviços e escopos, como mostrado na Figura 17.11.

FIGURE 17.10 Access Scopes section in VM instance detail edit page

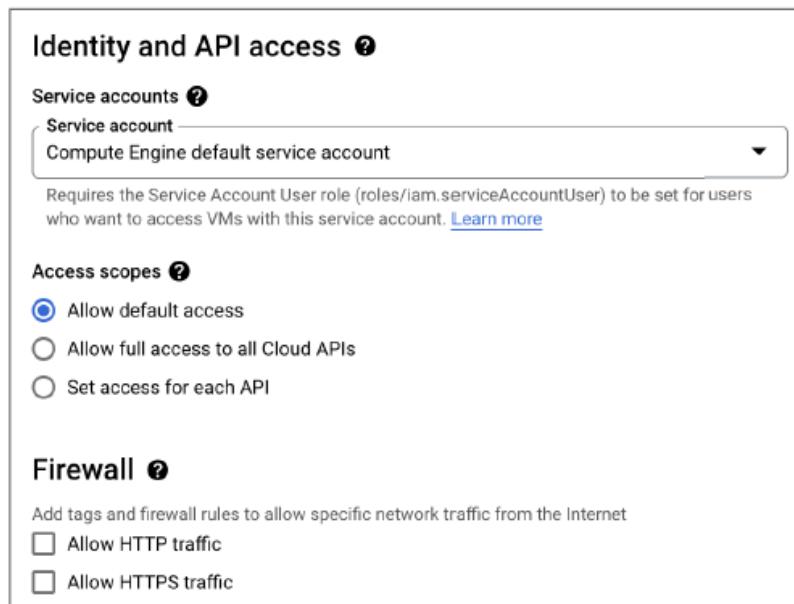
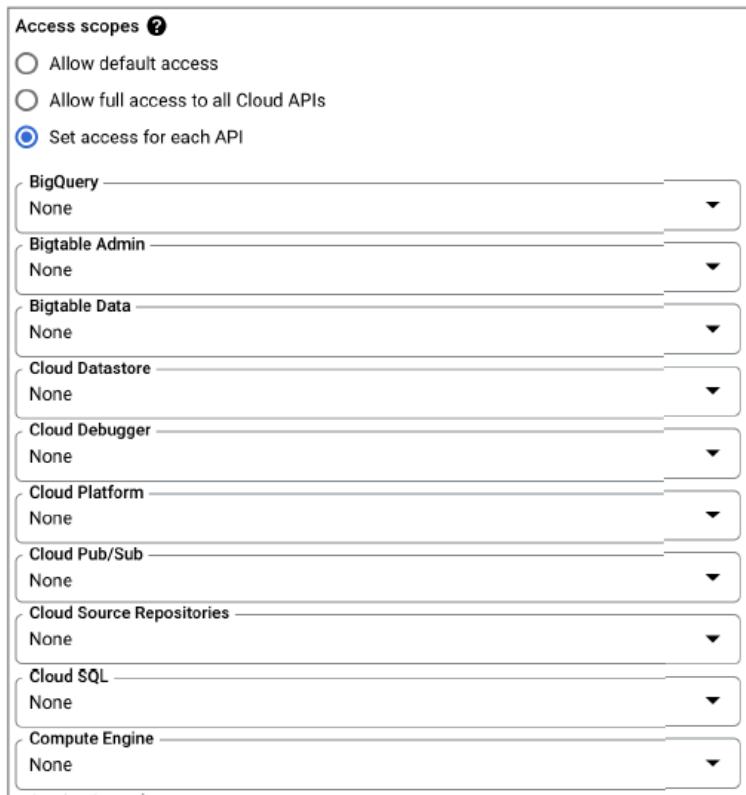


FIGURE 17.11 A partial list of services and scopes that can be individually configured



Você também pode definir escopos usando o comando `gcloud compute instances set-service-account`. A estrutura do comando é a seguinte:

```
gcloud compute instances set-service-account [NOME_DA_INSTÂNCIA] --service-account [EMAIL_DA_CONTA_DE_SERVIÇO] | --no-service-account --no-scopes | --scopes [ESCOPOS,...]
```

Um exemplo de atribuição de escopo usando gcloud é o seguinte:

```
gcloud compute instances set-service-account ace-instance --service-account examadmin@ace-exam-project.iam.gserviceaccount.com --scopes compute-rw,storage-ro
```

Atribuindo uma Conta de Serviço a uma Instância de VM

Você pode atribuir uma conta de serviço a uma instância de VM. Primeiro, crie uma conta de serviço navegando até a seção Contas de Serviço da seção IAM & Admin do console. Clique em Criar Conta de Serviço para exibir uma página como a mostrada na Figura 17.12.

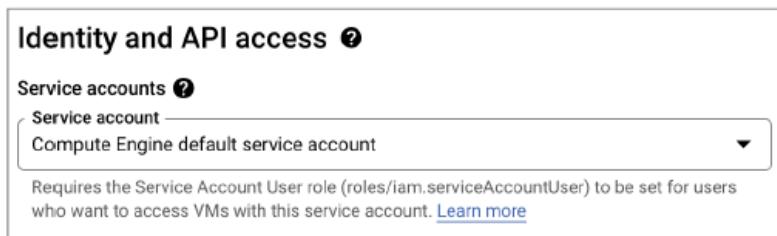
FIGURE 17.12 Creating a service account in the console

The screenshot shows the 'Create service account' dialog box. It has two main sections: 'Service account details' and 'Grant this service account access to project (optional)'. The 'Service account details' section contains fields for 'Service account name', 'Display name for this service account', 'Service account ID *' (with a delete and copy icon), 'Email address' (with a refresh icon), and 'Service account description'. Below these is a 'CREATE AND CONTINUE' button. The 'Grant this service account access to project (optional)' section has three numbered steps: 1. 'Grant this service account access to project (optional)', 2. 'Grant users access to this service account (optional)', and 3. 'Grant this service account access to specific APIs (optional)'. At the bottom are 'DONE' and 'CANCEL' buttons.

Após especificar um nome, identificador e descrição, clique em Criar e continuar. Em seguida, você pode atribuir funções conforme descrito anteriormente, usando o console ou comandos gcloud. Uma vez que tenha atribuído as funções que deseja que a conta de serviço tenha, você pode atribuí-la a uma instância de VM.

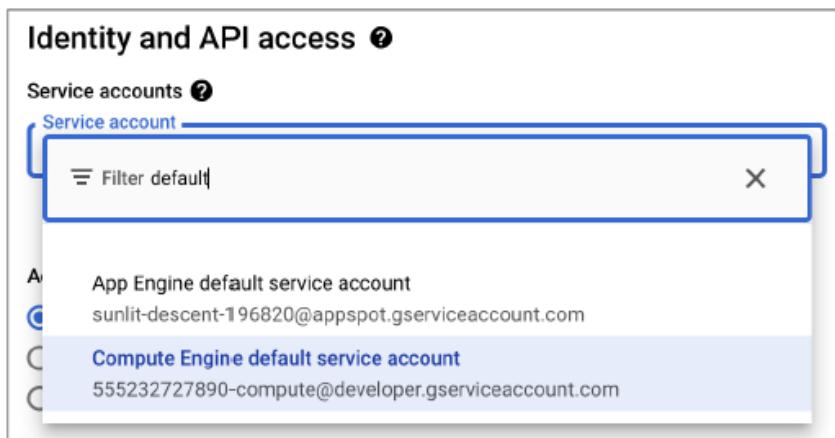
Navegue até a página Instâncias de VM na seção Compute Engine do console. Selecione uma instância de VM e clique em Editar. Isso exibirá uma página com um parâmetro para a instância. Role para baixo para ver o parâmetro rotulado Conta de Serviço (veja a Figura 17.13)

FIGURE 17.13 Section of Edit Instance page showing the Service Account parameter



Da lista suspensa, selecione a conta de serviço que deseja atribuir a essa instância, conforme mostrado na Figura 17.14.

FIGURE 17.14 List of service accounts that can be assigned to the instance



Você também pode especificar uma instância de serviço na linha de comando ao criar uma instância usando o comando `gcloud compute instances create`. Ele tem a seguinte estrutura:

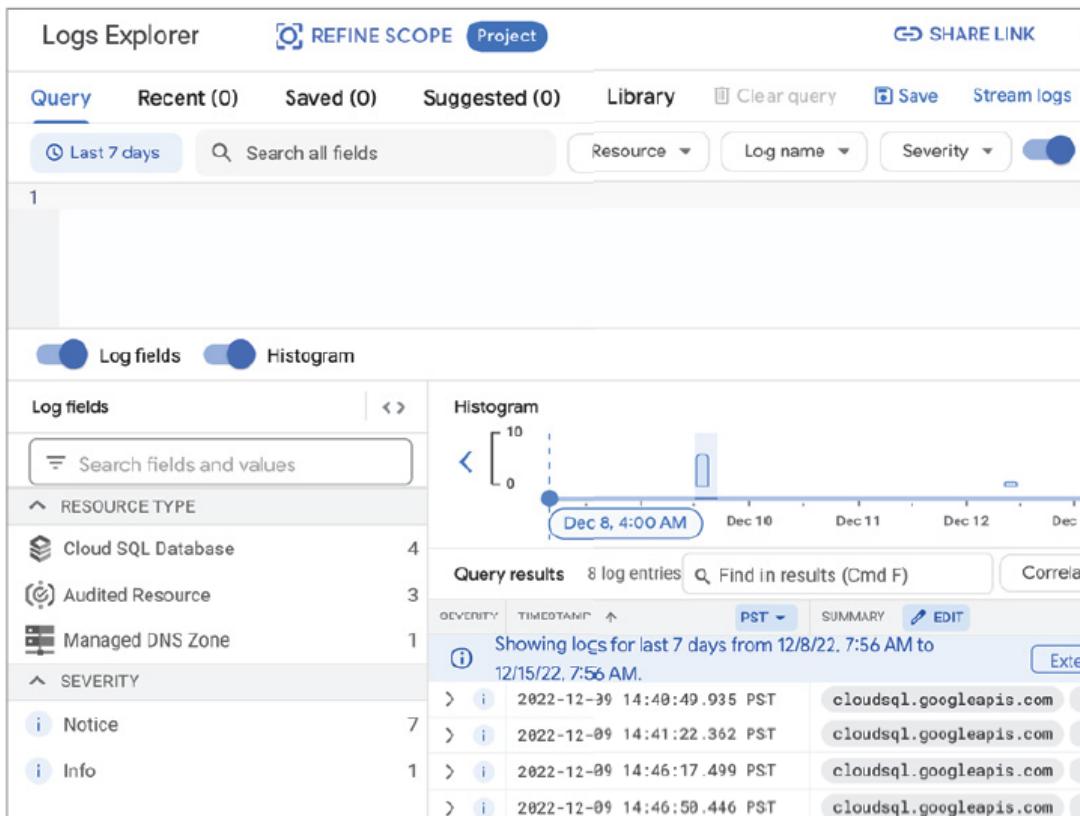
```
gcloud compute instances create [NOME_DA_INSTÂNCIA] --service-account  
[EMAIL_DA_CONTA_DE_SERVIÇO]
```

Para conceder acesso a um projeto, navegue até a página IAM do console e adicione um membro. Use o email da conta de serviço como a entidade a adicionar.

Visualizando Logs de Auditoria

Para visualizar logs de auditoria, navegue até a página Cloud Logging no Cloud Console.

FIGURE 17.15 Default listing of the Cloud Logging page



Isso mostrará uma listagem como a mostrada na Figura 17.15. Observe que você pode selecionar o recurso, tipos de logs para exibir, o nível do log e o período de tempo do qual exibir entidades. Usando o Nome do Log, procure por atividade para ver logs de auditoria de Atividade, data_access para ver logs de auditoria de Acesso a Dados, system_event para ver logs de auditoria de Evento de Sistema, e policy para ver logs de auditoria de Política Negada.

Para informações adicionais sobre logging, veja o Capítulo 18, “Monitoramento, Logging e Estimativa de Custos”.

Resumo

Os controles de acesso no Google Cloud são gerenciados usando IAM, funções básicas e escopos. As três funções básicas são Proprietário, Editor e Visualizador. Elas fornecem controles de acesso grosseiros aos recursos. Escopos são controles de acesso que se aplicam a instâncias de VMs. Eles são usados para limitar operações que podem ser realizadas por uma instância. O conjunto de operações que uma instância pode realizar é determinado pelos escopos atribuídos e pelas funções atribuídas a uma conta de serviço usada pela instância. O IAM fornece funções predefinidas. Essas funções são agrupadas por serviço. As funções são projetadas para fornecer o conjunto mínimo de permissões necessárias para realizar uma tarefa lógica, como escrever em um bucket ou implantar um aplicativo no App Engine. Quando as funções predefinidas não atendem às suas necessidades, você pode definir funções personalizadas.

Contas de serviço são usadas para permitir que VMs realizem operações com um conjunto de permissões. As permissões são concedidas às contas de serviço através das funções atribuídas à conta de serviço. Você pode usar a conta de serviço padrão fornecida pelo Google Cloud para uma instância ou pode atribuir a sua própria.

Essenciais para o Exame

Conheça os três tipos de funções: básicas, predefinidas e personalizadas. Funções básicas incluem Proprietário, Editor e Visualizador. Estas foram desenvolvidas antes do lançamento do IAM. Funções predefinidas são funções do IAM. Permissões são atribuídas a essas funções, e então as funções são atribuídas a usuários, grupos e contas de serviço. Funções personalizadas incluem permissões selecionadas pelo usuário que cria a função personalizada.

Entenda que escopos são um tipo de controle de acesso aplicado a instâncias de VM. A VM só pode realizar operações permitidas pelos escopos e funções do IAM atribuídas à conta de serviço da instância. Você pode usar funções do IAM para restringir escopos e usar escopos para restringir funções do IAM.

Saiba como visualizar funções atribuídas a identidades. Você pode usar a aba Funções na seção IAM & Admin do console para listar as identidades atribuídas a funções particulares. Você também pode usar o comando gcloud projects get-iam-policy para listar funções atribuídas a usuários em um projeto.

Entenda que funções do IAM suportam separação de funções e o princípio do menor privilégio. Funções básicas não suportavam o menor privilégio e separação de funções porque são muito genéricas. A separação de funções garante que duas ou mais pessoas sejam necessárias para completar uma tarefa sensível.

Saiba como usar gcloud iam roles describe para visualizar detalhes de uma função, incluindo permissões atribuídas a uma função. Você também pode visualizar as funções concedidas aos usuários aprofundando-se em uma função na página de Funções da seção IAM & Admin do console. Ao trabalhar com IAM, você estará usando o comando gcloud quando estiver trabalhando a partir da linha de comando.

Entenda as diferentes opções para acessar escopos ao criar uma instância. As opções são Acesso Padrão, Acesso Total e Definir Acesso Para Cada API. Se você não tiver certeza de qual usar, você pode conceder acesso total, mas certifique-se de limitar o que a instância pode fazer atribuindo funções que restringem as operações permitidas.

Saiba que o Cloud Logging coleta eventos de log. Eles podem ser filtrados e exibidos na seção Logging do Cloud Console. Você pode filtrar por recurso, tipo de log, nível de log e período de tempo para exibir.

Questões de Revisão

Você pode encontrar as respostas no Apêndice.

1. O que IAM significa?
 - A. Gerenciamento de identidade e autorização
 - B. Gerenciamento de identidade e acesso
 - C. Gerenciamento de identidade e auditoria
 - D. Gerenciamento de acesso individual
2. Quando você navega para IAM & Admin no Console da Nuvem, o que aparece no corpo principal da página?
 - A. Membros e funções atribuídas
 - B. Apenas funções
 - C. Apenas membros
 - D. Funções e permissões atribuídas
3. Por que as funções básicas são classificadas em uma categoria além do IAM?
 - A. Elas fazem parte do IAM.
 - B. Elas foram criadas antes do IAM.
 - C. Elas foram criadas após o IAM.
 - D. Elas não estão relacionadas ao controle de acesso.
4. Um estagiário desenvolvedor está confuso sobre para que servem as funções. Você descreve as funções do IAM como uma coleção de quê?
 - A. Identidades
 - B. Permissões
 - C. Listas de controle de acesso
 - D. Registros de auditoria
5. Você deseja listar as funções atribuídas aos usuários em um projeto chamado ace-exam-project. Qual comando gcloud você usaria?
 - A. gcloud iam get-iam-policy ace-exam-project
 - B. gcloud projects list ace-exam-project
 - C. gcloud projects get-iam-policy ace-exam-project
 - D. gcloud iam list ace-exam-project

6. Você está trabalhando no formulário exibido após clicar no link Adicionar na página IAM do IAM & Admin no Console da Nuvem. Existe um campo chamado Novos Membros. Que itens você inseriria nesse parâmetro?
- A. Apenas usuários individuais
 - B. Usuários individuais ou grupos
 - C. Funções ou usuários individuais
 - D. Funções ou grupos
7. Você foi atribuído a função de Publicador do App Engine. Que operações você pode realizar?
- A. Escrever novas versões de uma aplicação apenas.
 - B. Ler a configuração e as configurações da aplicação apenas.
 - C. Ler a configuração e as configurações da aplicação e escrever novas configurações.
 - D. Ler a configuração e as configurações da aplicação e escrever novas versões.
8. Você deseja listar permissões em uma função usando o Console da Nuvem. Onde você iria para ver isso?
- A. IAM & Admin; selecione Funções. Todas as permissões serão exibidas.
 - B. IAM & Admin; selecione Funções. Marque a caixa ao lado de uma função para exibir as permissões nessa função.
 - C. IAM & Admin; selecione Registros de Auditoria.
 - D. IAM & Admin; selecione Contas de Serviço e depois Funções.
9. Você está se reunindo com um auditor para discutir práticas de segurança na nuvem. O auditor pergunta como você implementa várias melhores práticas. Você descreve como as funções predefinidas do IAM ajudam a implementar qual(is) melhor(es) prática(s) de segurança?
- A. Privilégio mínimo
 - B. Separação de deveres
 - C. Defesa em profundidade
 - D. Opções A e B
10. Quais estágios de lançamento estão disponíveis ao criar funções personalizadas?
- A. Apenas alfa e beta
 - B. Apenas disponibilidade geral
 - C. Apenas desativado
 - D. Alfa, beta, disponibilidade geral e desativado

11. Qual é o comando gcloud usado para criar uma função personalizada?
- A. gcloud project roles create
 - B. gcloud iam roles create
 - C. gcloud project create roles
 - D. gcloud iam create roles
12. Um engenheiro de DevOps está confuso sobre o propósito dos escopos. Escopos são controles de acesso que são aplicados a que tipo de recursos?
- A. Buckets de armazenamento
 - B. Instâncias VM
 - C. Discos persistentes
 - D. Sub-redes
13. Um escopo é identificado usando que tipo de identificador?
- A. Um ID gerado aleatoriamente
 - B. Uma URL que começa com www.googleapisaccounts
 - C. Uma URL que começa com www.googleapis.com/auth
 - D. Uma URL que começa com www.googleapis.com/auth/PROJECT_ID
14. Uma instância VM está tentando ler de um bucket do Cloud Storage. A leitura do bucket é permitida pelas funções do IAM concedidas à conta de serviço da VM. A leitura dos buckets é negada pelos escopos atribuídos à VM. O que acontecerá se a VM tentar ler do bucket?
- A. O aplicativo que realiza a leitura ignorará a operação de leitura.
 - B. A leitura será executada porque a permissão mais permissiva é permitida.
 - C. A leitura não será executada porque tanto os escopos quanto as funções do IAM são aplicados para determinar quais operações podem ser realizadas.
 - D. A operação de leitura terá sucesso, mas uma mensagem será registrada no Cloud Logging.
15. Quais são as opções para definir escopos em uma VM?
- A. Permitir Acesso Padrão e Permitir Acesso Total apenas.
 - B. Permitir Acesso Padrão, Permitir Acesso Total e Definir Acesso Para Cada API.
 - C. Permitir Acesso Total ou Definir Acesso Para Cada API Apenas.
 - D. Permitir Acesso Padrão e Definir Acesso Para Cada API Apenas.
16. Qual comando gcloud você usaria para definir escopos?
- A. gcloud compute instances set-scopes

- B. gcloud compute instances set-service-account
 - C. gcloud compute service-accounts set-scopes
 - D. gcloud compute service-accounts define-scopes
17. Qual comando gcloud você usaria para atribuir uma conta de serviço ao criar uma VM?
- A. gcloud compute instances create [NOME_DA_INSTÂNCIA]
--service-account [EMAIL_DA_CONTA_DE_SERVIÇO]
 - B. gcloud compute instances create-service-account [NOME_DA_INSTÂNCIA][EMAIL_DA_CONTA_DE_SERVIÇO]
 - C. gcloud compute instances define-service-account [NOME_DA_INSTÂNCIA][EMAIL_DA_CONTA_DE_SERVIÇO]
 - D. gcloud compute instances-service-account create [NOME_DA_INSTÂNCIA][EMAIL_DA_CONTA_DE_SERVIÇO]
18. Qual serviço do Google Cloud é usado para visualizar registros de auditoria?
- A. Compute Engine
 - B. Cloud Storage
 - C. Cloud Logging
 - D. Log personalizado
19. Quais opções estão disponíveis para filtrar mensagens de log ao visualizar registros de auditoria?
- A. Período de tempo e nível de log apenas
 - B. Recurso, tipo de log, nível de log e período de tempo apenas
 - C. Recurso e período de tempo apenas
 - D. Tipo de log apenas
20. Um auditor precisa revisar os registros de auditoria. Você atribui permissão somente leitura a uma função personalizada que você cria para auditores. Que prática de segurança você está seguindo?
- A. Defesa em profundidade
 - B. Privilégio mínimo
 - C. Separação de deveres
 - D. Varredura de vulnerabilidade

Capítulo 18

Monitoramento, Registro e Estimativa de Custos

ESTE CAPÍTULO COBRE OS SEGUINtes OBJETIVOS DO EXAME DE CERTIFICAÇÃO GOOGLE ASSOCIATE CLOUD ENGINEER:

- ✓✓ 2.1 Planejamento e estimativa do uso dos produtos Google Cloud usando a Calculadora de Preços
- ✓✓ 4.6 Monitoramento e registro

Monitorar o desempenho do sistema é uma parte essencial da engenharia de nuvem. Neste capítulo, você aprenderá sobre o conjunto de Operações na Nuvem, um serviço do Google Cloud para monitoramento de recursos, registro e rastreamento. Você começará criando alertas baseados em métricas de recursos e métricas personalizadas. Em seguida, você voltará sua atenção para o registro, com uma discussão sobre como criar sinks de log para armazenar dados de registro fora das Operações na Nuvem. Você também verá como visualizar e filtrar dados de log. Operações na Nuvem inclui ferramentas de diagnóstico, como o Cloud Trace, sobre o qual você aprenderá também. Concluiremos o capítulo com uma revisão da Calculadora de Preços para estimar o custo dos recursos e serviços do Google Cloud.

Monitoramento na Nuvem

O Monitoramento na Nuvem é um serviço para coletar métricas de desempenho, registros e dados de eventos de nossos recursos. Métricas incluem medições como a porcentagem média de utilização da CPU no último minuto e o número de bytes escritos em um dispositivo de armazenamento no último minuto. O Monitoramento na Nuvem inclui muitas métricas predefinidas. Alguns exemplos são mostrados na Tabela 18.1 que você pode usar para avaliar a saúde dos seus recursos e, se necessário, disparar alertas para chamar sua atenção para recursos ou serviços que não estão atendendo aos objetivos de nível de serviço.

TABLE 18.1 Example Cloud Monitoring metrics

Google Cloud Product	Metric
Compute Engine	CPU utilization
Compute Engine	Disk bytes read
BigQuery	Execution times
Bigtable	CPU load
Cloud Functions	Execution count

O Monitoramento na Nuvem funciona em ambientes híbridos, com suporte para Google Cloud, Amazon Web Services e recursos locais.

Criando Painéis

Métricas são medições definidas em um recurso coletadas em intervalos regulares. Métricas retornam valores agregados, como o valor máximo, mínimo ou médio do item medido, que pode ser a utilização da CPU, quantidade de memória usada ou número de bytes escritos em uma interface de rede.

Para este exemplo, assuma que você está trabalhando com uma VM que tem o Apache Server e PHP instalados. VMs coletarão métricas e registros básicos, mas para métricas mais detalhadas, você pode instalar o Ops Agent, que inclui suporte tanto para monitoramento quanto para registro. Para instalar o Ops Agent em uma VM Linux, execute o seguinte comando no prompt de comando (note que estes não são comandos gcloud):

```
curl -sSO https://dl.google.com/cloudagents/add-google-cloud-ops-agent-repo.sh  
sudo bash add-google-cloud-ops-agent-repo.sh --also-install
```

VMs com agentes instalados coletam dados de monitoramento e registro e os enviam para o Monitoramento na Nuvem e Registro na Nuvem.

A Figura 18.1 mostra um exemplo da página de Visão Geral do Monitoramento na Nuvem. Ela inclui informações sobre o status da configuração de monitoramento, painéis disponíveis, bem como links para artigos relacionados e postagens em blogs.

Detalhes sobre incidentes e verificações de saúde também estão disponíveis na visão geral.

FIGURE 18.1 Partial view of Cloud Monitoring Overview page

The screenshot shows the 'Overview' section of the Cloud Monitoring interface. On the left, there's a sidebar with various icons representing different monitoring categories. The main area has two main sections: 'Dashboards' and 'Favorites'. The 'Dashboards' section includes a 'Create Dashboard' button, a 'Infrastructure' section with links to VM Instances, GKE, BigQuery, Cloud Storage, and App Engine, and a link to 'View all resource dashboards'. The 'Favorites' section shows a globe icon with colored dots (red, blue, green) and a link to 'View your favorite dashboards'. At the bottom, there are buttons for 'CREATE POLICY', 'SHOW CLOSED INCIDENTS', and a dropdown menu.

No painel esquerdo da página de Monitoramento na Nuvem, você pode selecionar outras visualizações de monitoramento, incluindo painéis. Uma lista de exemplo de painéis padrão é mostrada na Figura 18.2. A lista inclui painéis para App Engine, BigQuery, Cloud Storage, Firewalls e VPN.

FIGURE 18.2 Available dashboards in Cloud Monitoring

The screenshot shows the 'Dashboards Overview' page. It features a 'DASHBOARD LIST' tab selected, showing a list of dashboards categorized by type: All, Recently Viewed, Favorites, Custom, GCP, Integrations, and Other. There are also 'Labeled' and 'SAMPLE LIBRARY' tabs. The main table lists dashboards with columns for Name, Type, and Labels. The dashboards listed are App Engine, BigQuery, Cloud Storage, Firewalls, and VPN, all categorized under 'Google Cloud Platform'.

Categories	All Dashboards	Manage Labels
Filter by category	Filter Dashboards	
All	App Engine	Google Cloud Platform
Recently Viewed	BigQuery	Google Cloud Platform
Favorites	Cloud Storage	Google Cloud Platform
Custom	Firewalls	Google Cloud Platform
GCP	VPN	Google Cloud Platform
Integrations		
Other		
Labeled		

Cada um dos painéis mostra informações relevantes para o serviço. Por exemplo, o painel do Cloud Storage mostra dados sobre incidentes, buckets, solicitações e tráfego de rede enviado, conforme mostrado na Figura 18.3.

Além dos painéis predefinidos disponíveis no Monitoramento na Nuvem, você pode criar os seus próprios clicando em Criar Painel para exibir a janela mostrada na Figura 18.4.

Se você escolher criar um gráfico de linhas, um componente de gráfico é adicionado ao painel, conforme mostrado na Figura 18.5. Neste exemplo, será traçada a métrica de utilização média da CPU.

Usando o Explorador de Métricas

O Explorador de Métricas é outra funcionalidade do Monitoramento na Nuvem. Ele permite que você visualize uma ampla variedade de métricas escolhendo de uma lista de métricas. A Figura 18.6 mostra a página principal do Explorador de Métricas, e a Figura 18.7 mostra detalhes das métricas que você pode visualizar relacionadas aos buckets do Cloud Storage.

Depois de selecionar a métrica de contagem de objetos para os buckets do Cloud Storage, o Explorador de Métricas exibe um gráfico de linhas, conforme mostrado na Figura 18.8.

FIGURE 18.3 Cloud Storage monitoring dashboard

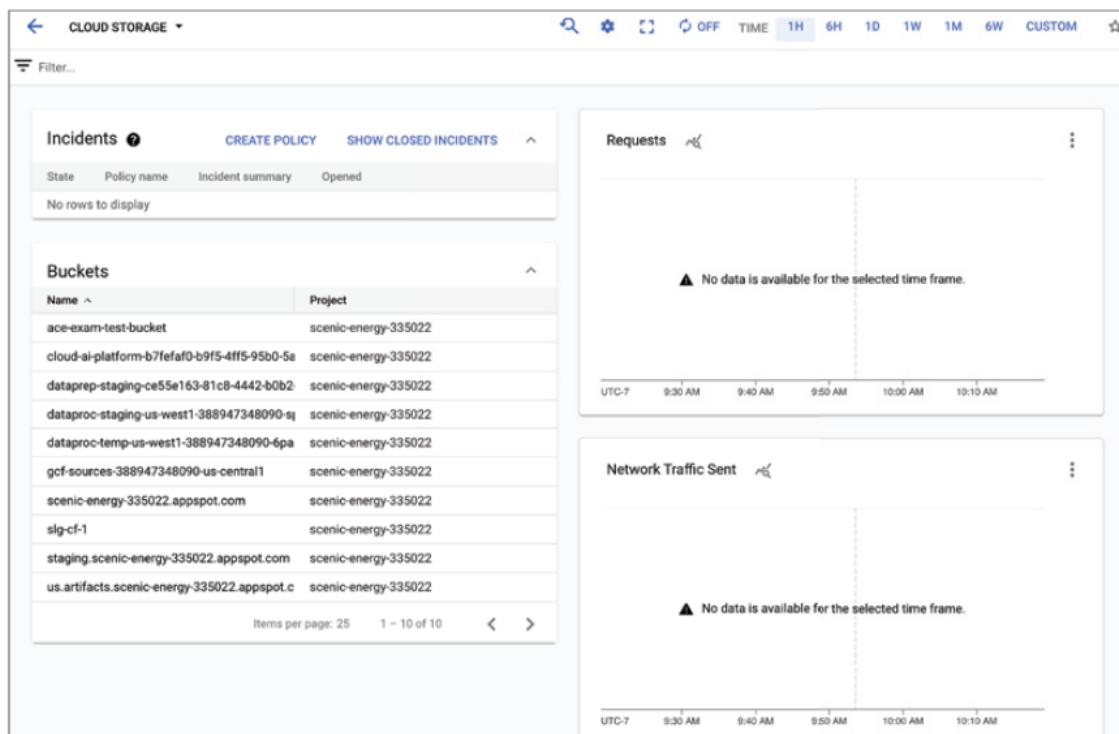


FIGURE 18.4 Creating your own dashboard begins with choosing a chart.

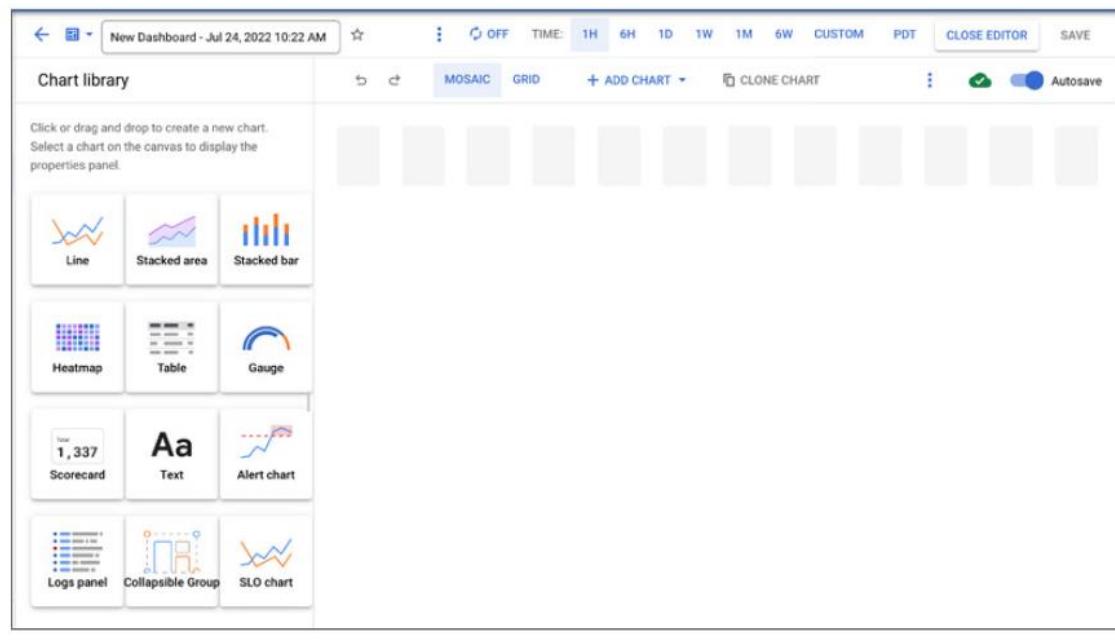


FIGURE 18.5 Adding a line chart to display mean CPU utilization

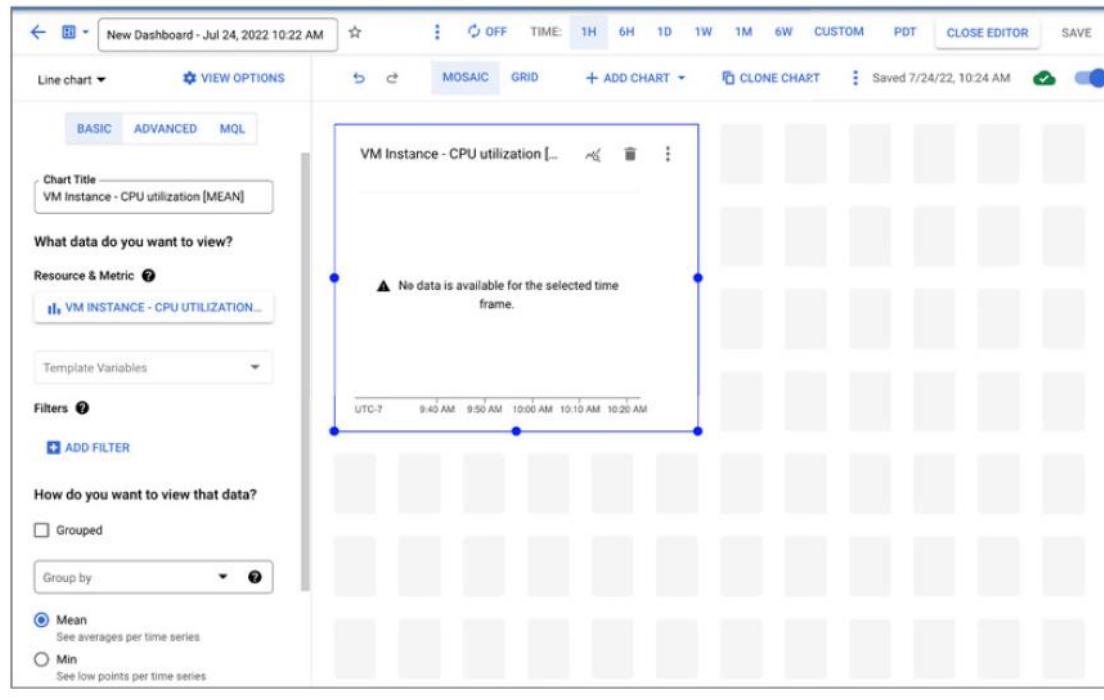


FIGURE 18.6 Main page of Metric Explorer

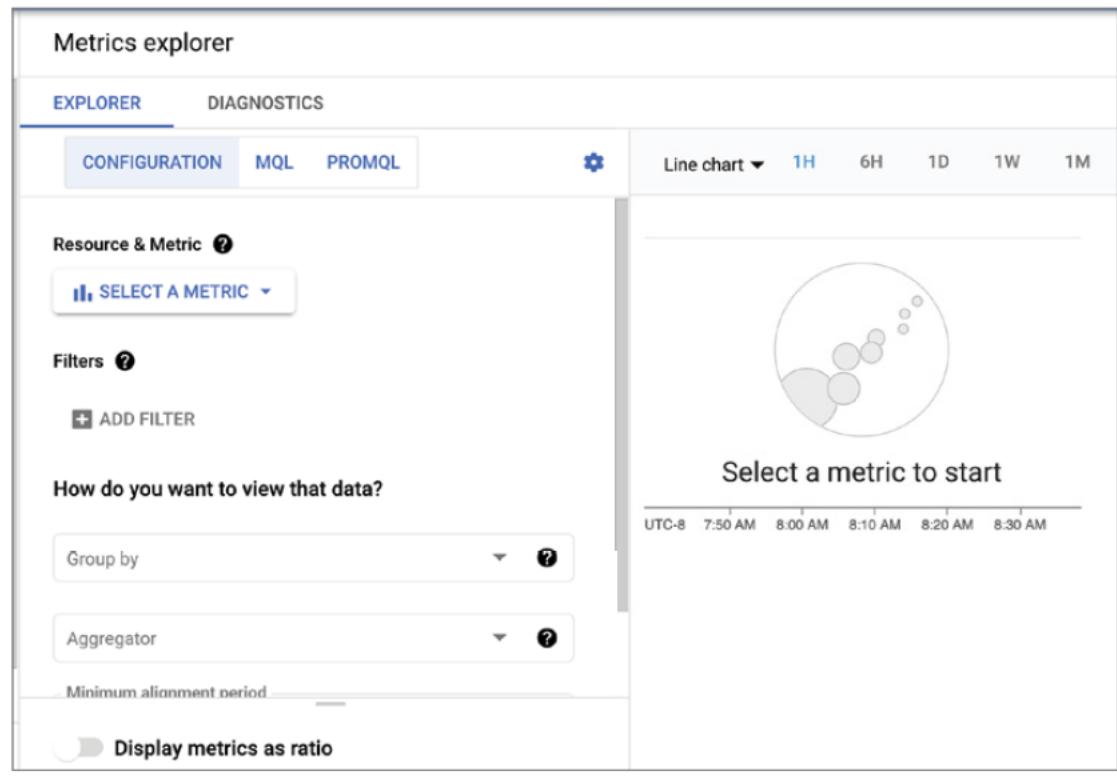


FIGURE 18.7 Metrics available for Cloud Storage Buckets

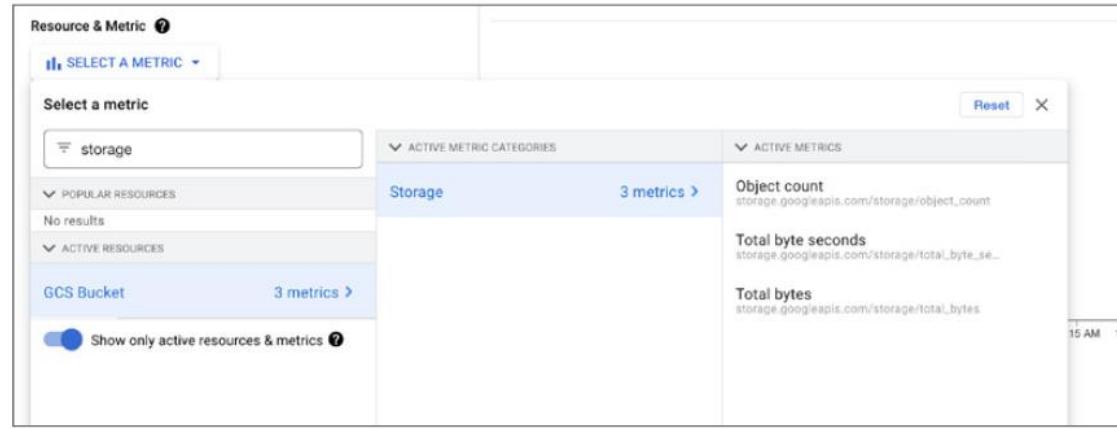
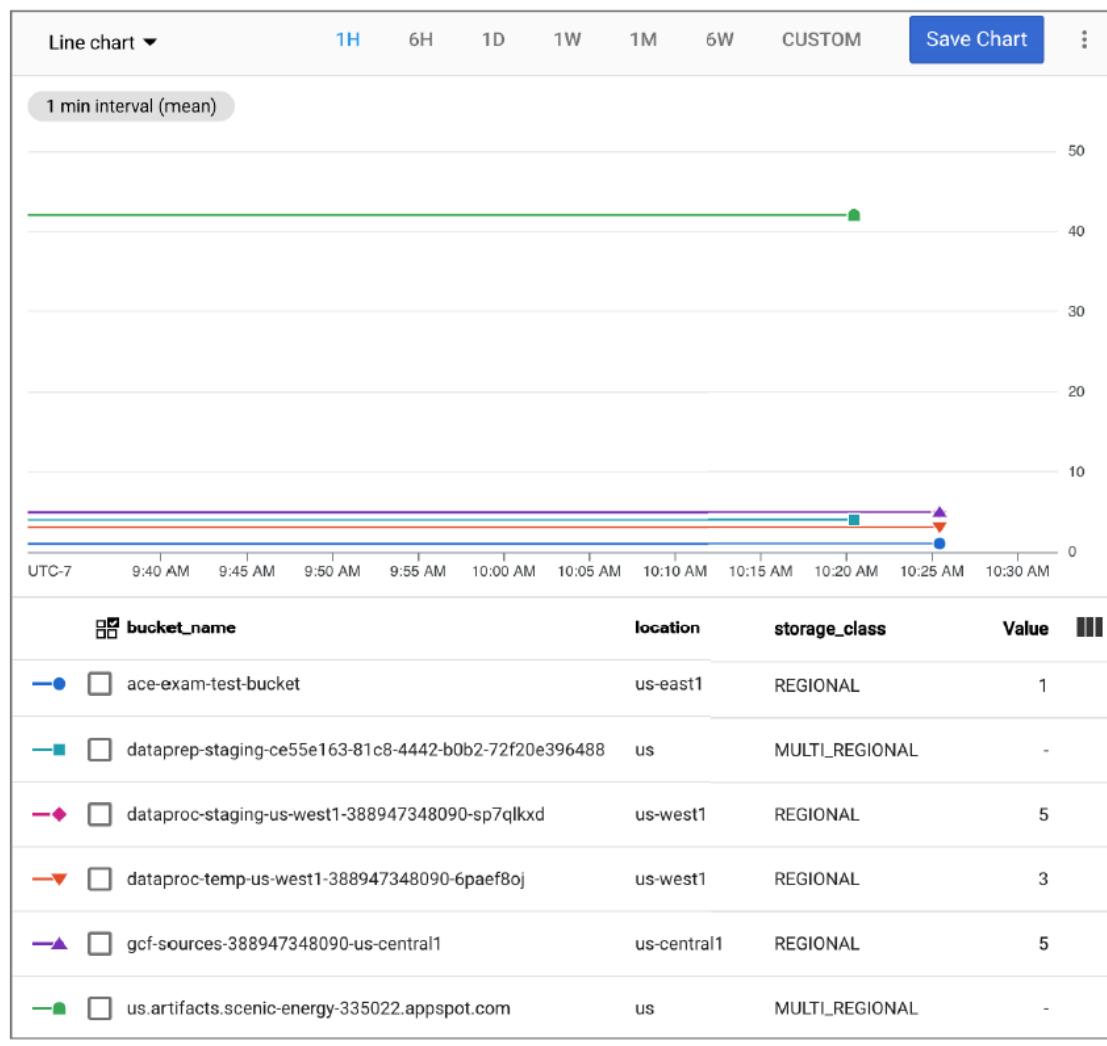


FIGURE 18.8 Line chart of object count metric for Cloud Storage buckets



Criando Alertas

Os painéis são úteis para obter uma visão rápida de um conjunto de métricas-chave, e o Explorador de Métricas é útil quando você está investigando um problema e precisa visualizar uma variedade de métricas.

Se você quiser ser notificado automaticamente quando uma métrica exceder algum limite, você pode criar alertas.

A página principal de Alertas no Monitoramento na Nuvem é mostrada na Figura 18.9. Ela inclui um resumo da contagem de incidentes ativos, incidentes respondidos e políticas de alerta. Também há listagens detalhadas de incidentes e políticas.

FIGURE 18.9 Alerting main page of Cloud Logging

The screenshot shows the 'Alerting' main page in the Google Cloud Platform. At the top, there are two buttons: '+ CREATE POLICY' and 'EDIT NOTIFICATION CHANNELS'. Below these, a message states: 'Monitoring now supports both user-scoped and device-scoped Cloud Console Mobile notification channels'. There are three links: 'MANAGE CHANNELS', 'LEARN MORE', and 'DISMISS'. A summary section follows, showing 'Incidents firing' (0), 'Incidents acknowledged' (0), and 'Alert policies' (0). A link 'View all' is provided for alert policies. The next section is titled 'Incidents' with a 'SHOW CLOSED INCIDENTS' button. It includes columns for 'State', 'Policy name', 'Incident summary', and 'Opened'. A message says 'No rows to display' and has a link '→ See all incidents'. The final section is titled 'Snoozes' with a 'PREVIEW' and 'CREATE SNOOZE' button. It includes columns for 'State', 'Name', 'Start time', and 'End time'. A link 'Show past snoozes' is also present. A message 'No rows to display' is shown here as well.

Para criar um alerta, você cria uma política. Uma política é definida para uma métrica. Por exemplo, a Figura 18.10 mostra o início da definição de uma política para alertá-lo quando houver um acúmulo de mensagens não reconhecidas em um tópico do Cloud Pub/Sub.

FIGURE 18.10 Creating a policy for a Pub/Sub backlog

Select a metric [?](#)

CLOUD PUB/SUB SUBSCRIPTION - BACKLOG SIZE

Add filters Optional

Selections made on the chart do not affect the alert policy [Learn more](#)

ADD FILTER

Transform data

Within each time series [?](#)

Rolling window * [5 min](#)

Adjust the length of time a signal is calculated for. Example: Mean of CPU utilization for 5 minutes is above 80%

Rolling window function * [mean](#)

Function applied to the rolling window

Across time series

Add secondary data transformation

NEXT

Para uma política, você também especifica o tipo de condição, que pode ser um limite ou uma ausência de métrica. Uma condição de limite é acionada quando o valor da métrica está acima ou cai abaixo do valor especificado pelo período de tempo especificado. Uma condição de ausência de métrica é baseada na ausência de dados por um período de tempo especificado (veja a Figura 18.11).

Você também especifica um gatilho de alerta, que especifica o escopo dos dados que você considera ao verificar a condição do alerta. A Figura 18.12 mostra as opções possíveis, que incluem Qualquer Série Temporal Viola, Percentual de Série Temporal Viola, Número de Série Temporal Viola e Todas as Séries Temporais Violam.

O último passo da criação de uma política é especificar os canais de notificação, conforme mostrado na Figura 18.13.

As opções para canais de notificação incluem:

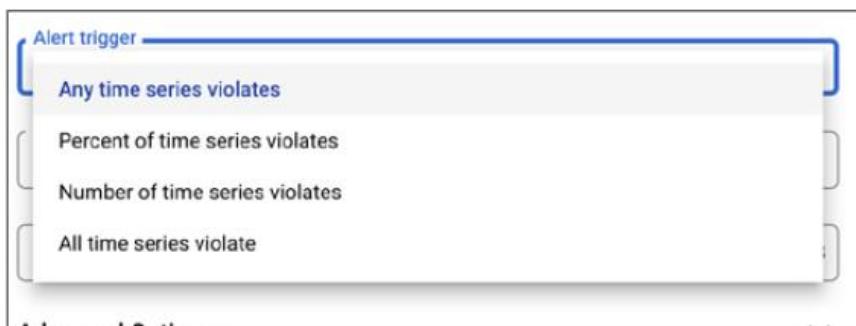
- Email, que envia mensagens para um endereço de email
- Slack, que envia mensagens para canais do Slack
- SMS, que envia mensagens de texto
- Cloud Pub/Sub, que posta mensagens em um tópico do Cloud Pub/Sub

- PagerDuty, que envia mensagens para uma plataforma SaaS popular para DevOps
- Webhooks, que invoca uma função de callback baseada em HTTP para enviar mensagens para um aplicativo

FIGURE 18.11 Configuring an alert

The screenshot shows a configuration interface for an alert trigger. At the top, there are buttons for '+ ADD ALERT CONDITION' and 'DELETE ALERT CONDITION'. Below this, the title 'Configure alert trigger' is displayed. Under 'Condition type', the radio button for 'Threshold' is selected, with a descriptive text explaining it triggers if a time series rises above or falls below a value for a specific duration window. The 'Metric absence' option is also available but unselected. The 'Alert trigger' dropdown is set to 'Any time series violates'. The 'Threshold position' dropdown is set to 'Above threshold'. The 'Threshold value' field contains the letter 'B'. An 'Advanced Options' section is collapsed, showing a 'Condition name' field containing 'Cloud Pub/Sub Subscription - Backlog size'. At the bottom right are 'CREATE POLICY' and 'PROVIDE FEEDBACK' buttons.

FIGURE 18.12 Alert trigger options



Você também pode especificar um período para fechar automaticamente o alerta, etiquetas e documentação para ser incluída com o alerta.

FIGURE 18.13 Creating notification channels for an alert

Configure notifications and finalize alert

Configure notifications Recommended

Use notification channel

Notification Channels ▾

Tip We recommend that you create multiple notification channels for redundancy purposes. Google has no control of many of the delivery systems after we have passed the notification to that system. Additionally, a single Google service supports Cloud Console Mobile App, PagerDuty, Webhooks, and Slack. If you use one of these notification channels, then use email, SMS, or Pub/Sub as the redundant channel.

[LEARN MORE](#)

Notify on incident closure

Incident autoclose duration * ▾
7 days

If data is absent, select a duration after which Incident will automatically close.

Policy user labels

Policy user labels allow you to add your own labels to alert policies for organization. The labels are included in the notification and incident details. (Optional)

[+ ADD LABEL](#)

Documentation Optional

Enter any documentation you would like included with the notification. You can use markdown, variables, and channel-specific controls. Markdown formatting may not apply to all notification channels.

Text Field ?

Muitos Alertas São Tão Ruins Quanto Poucos

Tenha cuidado ao elaborar políticas de monitoramento. Você não quer submeter os engenheiros a tantos alertas que eles comecem a ignorá-los. Isso é às vezes chamado de fadiga de alerta. Políticas que são muito sensíveis gerarão alertas quando nenhuma intervenção humana é necessária. Por exemplo, a utilização da CPU pode regularmente atingir picos por breves períodos de tempo. Se este é um padrão normal para o seu ambiente, e não está afetando negativamente sua capacidade de cumprir os acordos de nível de serviço, então há pouca razão para alertar sobre eles. Projete políticas para identificar condições que realmente requerem a atenção de um engenheiro e não são propensas a se resolver por conta própria. Use limiares que sejam longos o suficiente para que condições não sejam acionadas em estados transitórios que não durarão muito. Muitas vezes, até o tempo de um engenheiro resolvê-la, a condição já não está mais acionando. Projetar políticas de monitoramento é algo como uma arte. Você deve assumir que precisará de múltiplas iterações para ajustar suas políticas para encontrar o equilíbrio certo de gerar apenas os tipos certos de alertas úteis sem também gerar alertas que não são úteis.

Registro na Nuvem

O Registro na Nuvem é um serviço para coletar, armazenar, filtrar e visualizar dados de log e eventos. O registro é um serviço gerenciado, então você não precisa configurar ou implantar servidores para usar o serviço.

As diretrizes do Exame de Engenharia de Nuvem Associado observam várias tarefas de registro que um engenheiro de nuvem deve conhecer:

- Configurando roteadores de log
- Configurando sinks de log
- Visualizando e filtrando logs
- Visualizando detalhes da mensagem

Vamos revisar cada um destes nesta seção.

Roteadores de Log e Sinks de Log

Dados de log são ingeridos pela API de Registro na Nuvem. A partir daí, mensagens de log são roteadas para um dos três tipos de sinks: o sink de log Requerido, o sink de log Padrão ou um sink de log definido pelo usuário.

Sinks são associados a um recurso do Google Cloud, como uma conta de cobrança, projeto, pasta ou organização. O Google cria um sink Requerido e um Padrão para cada conta de cobrança, projeto, pasta ou organização.

O Roteador de Log é um serviço que recebe mensagens de log e aplica filtros de inclusão e exclusão para determinar quais sinks de log devem receber a mensagem. O Roteador de Log suporta o uso de combinações de sinks para rotear logs para múltiplos locais de armazenamento.

Configurando Sinks de Log

O sink de log Requerido é usado para armazenar atividades administrativas, eventos do sistema e logs de transparência de acesso. Esses logs são armazenados por 400 dias, e essa duração não pode ser alterada.

O sink de log Padrão recebe mensagens de log que não são enviadas para o sink de log Requerido. Esses logs são armazenados por 30 dias por padrão, mas você pode alterar isso configurando uma política de retenção personalizada. Uma retenção de 30 dias é suficiente se você usa logs para diagnosticar problemas operacionais, mas raramente visualiza os logs após alguns dias. Sua organização pode precisar manter logs por mais tempo para cumprir regulamentações governamentais ou do setor. Você também pode querer analisar os logs para obter insights sobre o desempenho da aplicação. Para esses casos de uso, é melhor exportar dados de log para um sistema de armazenamento de longo prazo como o Cloud Storage ou BigQuery.

Você pode criar buckets de log definidos pelo usuário em um projeto. Isso permite que você direcione um subconjunto de mensagens de log para um bucket específico do Cloud Storage. Você pode configurar uma retenção personalizada em um bucket de log definido pelo usuário.

Além de armazenar mensagens de log, o Cloud Logging suporta métricas de log. São métricas baseadas no conteúdo das mensagens de log. Se uma mensagem de log atende a um padrão de métrica de log, essa mensagem é refletida na métrica do Cloud Monitoring associada ao padrão.

O Cloud Logging suporta vários destinos onde as mensagens podem ser roteadas:

- Cloud Storage, para armazenamento de longo prazo de logs em formato JSON
- BigQuery, para logs que serão analisados
- Cloud Pub/Sub, para mensagens JSON consumidas por integrações de terceiros
- Cloud Logging, para visualização e armazenamento por períodos configuráveis pelo usuário

Visualizando e Filtrando Logs

Para visualizar o conteúdo dos logs, navegue até a seção Cloud Logging do console para visualizar a página do Explorador de Logs, mostrada na Figura 18.14.

O Explorador de Logs permite que você visualize mensagens de log. Como os logs são frequentemente bastante grandes, é importante poder filtrar rapidamente as mensagens para apenas aquelas que lhe interessam. O Explorador de Logs permite que você filtre mensagens com base em:

- Tempo (veja a Figura 18.15)
- Tipo de recurso (veja a Figura 18.16)
- Severidade (veja a Figura 18.17)

■■ Consulta de log (veja a Figura 18.18)

FIGURE 18.14 Log Explorer page of the Cloud Logging console

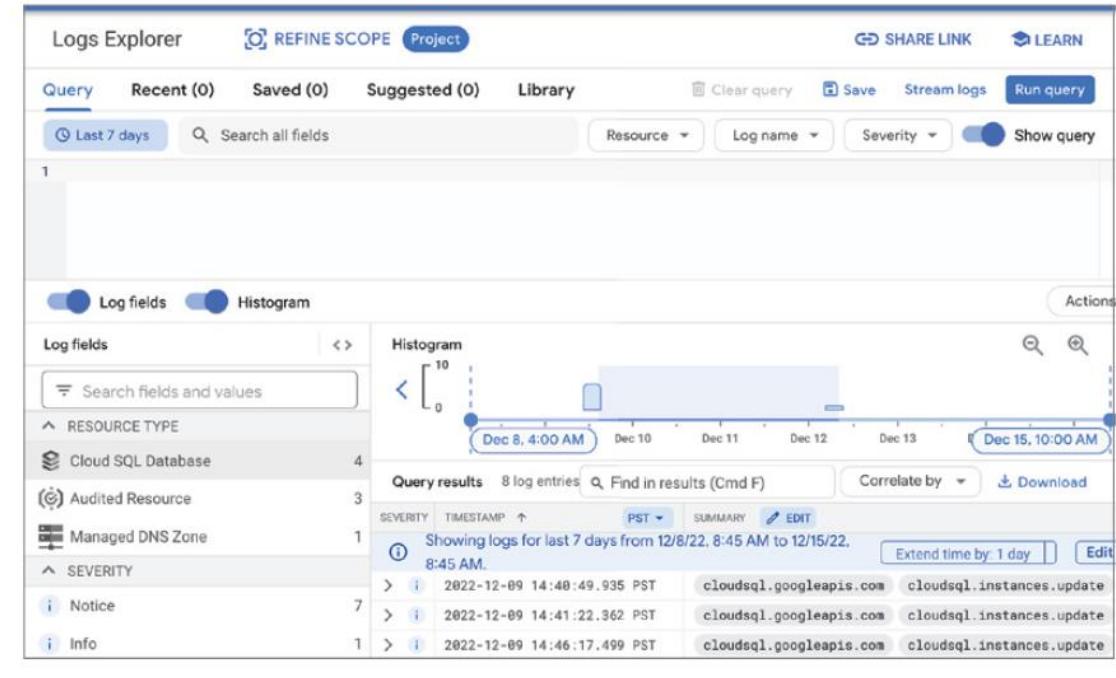


FIGURE 18.15 Time restriction options in Log Explorer

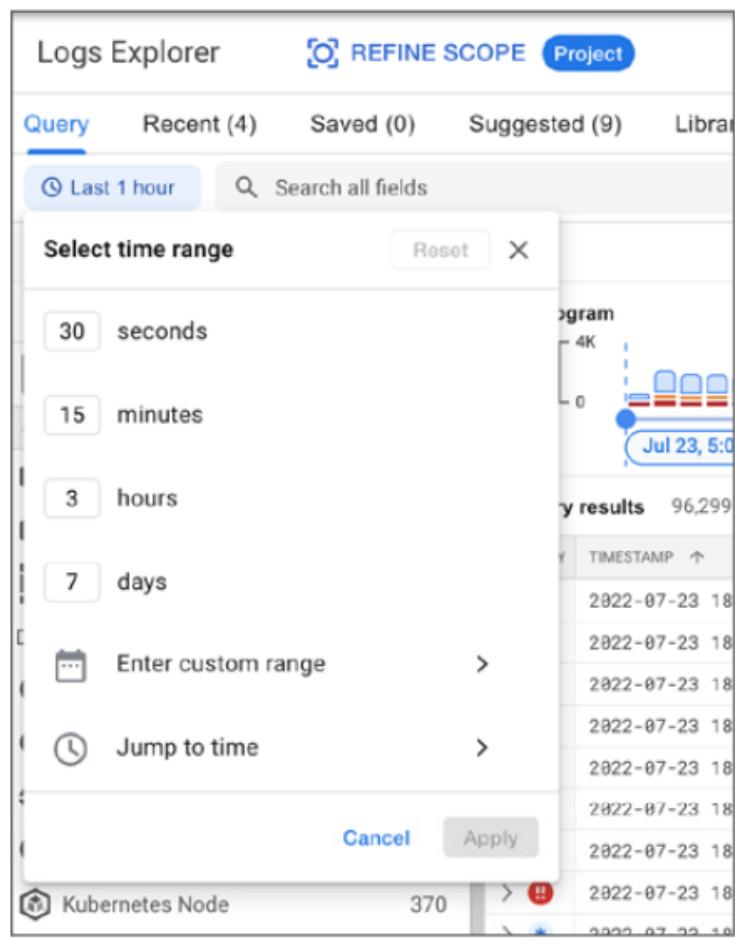


FIGURE 18.17 Severity filtering options in Log Explorer

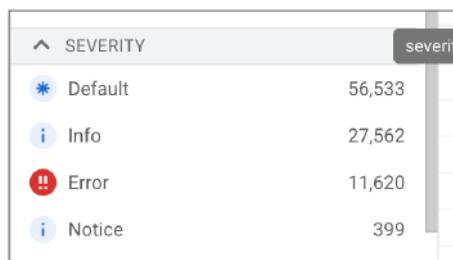
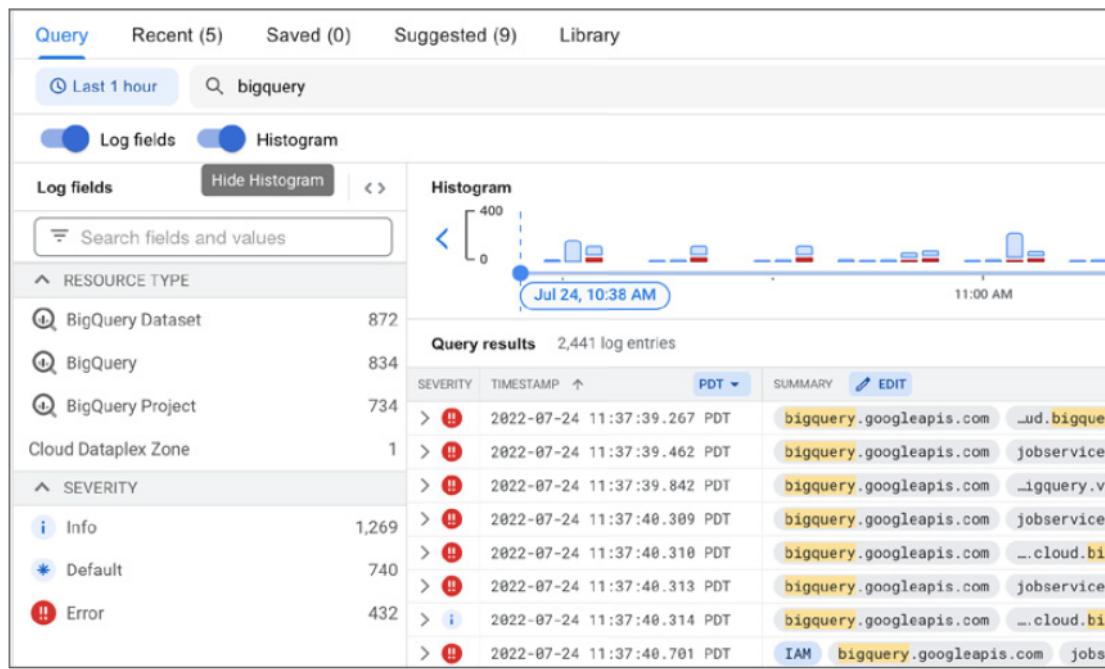


FIGURE 18.18 Queries in Log Explorer can be as simple as keyword searches.



Visualizando Detalhes da Mensagem

Cada entrada de log é exibida como uma única linha quando você visualiza o conteúdo dos logs. Repare no ícone de triângulo à esquerda da linha. Se você clicar nesse ícone, a linha se expandirá para mostrar detalhes adicionais. Por exemplo, a Figura 18.19 mostra uma entrada de log expandida em um nível.

FIGURE 18.19 A log entry expanded by one level

A screenshot of a log entry in Google Cloud Logging. The log entry is for a 'cloudsql.instances.update' event on '2022-12-09 14:40:49.935 PST'. The 'protoPayload' field is expanded, showing nested fields like 'insertId', 'logName', 'operation', 'protoPayload', 'resource', 'severity', and 'timestamp'. There are also buttons for 'Hide log summary', 'Expand nested fields', and 'Copy to clipboard'.

```
2022-12-09 14:40:49.935 PST    cloudsq...instances.update  
...c-energy-335022/instances/ace-exam-mysql  dan@sulliv...  
audit_log, method: "cloudsql.instances.update", princip...  
"dan@sullivanlearninggroup.com"  
{  
  insertId: "9zd91me941nf"  
  logName: "projects/scenic-energy-335022/logs/cloudaudit.googleapis.com%2Factivity"  
  operation: {3}  
  protoPayload: {10}  
  receiveTimestamp: "2022-12-09T22:40:50.570681137Z"  
  resource: {2}  
  severity: "NOTICE"  
  timestamp: "2022-12-09T22:40:49.935092Z"  
}
```

No caso da expansão de primeiro nível, você vê informações de alto nível como insertId, logName e receiveTimestamp. Você também vê outros elementos de dados estruturados, como protoPayload e recurso. A Figura 18.20 mostra a estrutura protoPayload expandida.

FIGURE 18.20 A log entry with the protoPayload structure expanded

A screenshot of a log entry in Google Cloud Logging, similar to Figure 18.19 but with the 'protoPayload' section expanded further. The expanded 'protoPayload' section shows detailed information about the audit log, including 'authenticationInfo', 'authorizationInfo', 'methodName', 'request', 'requestMetadata', 'resourceName', 'response', 'serviceName', and 'status'. There are also buttons for 'Hide log summary', 'Expand nested fields', and 'Copy to clipboard'.

```
{  
  insertId: "9zd91me941nf"  
  logName: "projects/scenic-energy-335022/logs/cloudaudit.googleapis.com%2Factivity"  
  operation: {3}  
  protoPayload: {  
    @type: "type.googleapis.com/google.cloud.audit.AuditLog"  
    authenticationInfo: {2}  
    authorizationInfo: [1]  
    methodName: "cloudsql.instances.update"  
    request: {4}  
    requestMetadata: {3}  
    resourceName: "projects/scenic-energy-335022/instances/ace-exam-mysql"  
    response: {12}  
    serviceName: "cloudsql.googleapis.com"  
    status: {0}  
  }  
}
```

Você pode continuar a explorar individualmente cada estrutura se houver um triângulo à esquerda. Por exemplo, na estrutura protoPayload, você poderia explorar mais a fundo em authenticationInfo, authorizationInfo e requestMetadata, entre outros.

A Figura 18.21 mostra a seção requestMetadata expandida.

FIGURE 18.21 Details of the requestMetadata section of a log message

The screenshot shows the Google Cloud Logging interface with the title 'Query results 2,441 log entries'. Below the title are filters for 'SEVERITY' (set to 'INFO'), 'TIMESTAMP' (sorted '↑'), 'PDT' (selected), 'SUMMARY' (selected), and 'EDIT' (button). The main area displays a single log entry with its 'protoPayload' field expanded. The expanded 'protoPayload' object includes fields like '@type', 'authenticationInfo', 'authorizationInfo', 'metadata', 'methodName', 'requestMetadata' (which further expands to show 'callerIp', 'callerSuppliedUserAgent', 'resourceName', 'serviceName', 'status', 'receiveTimestamp', 'resource', 'severity', and 'timestamp'). The 'methodName' field is highlighted in yellow, and the 'resourceName' field contains the URL 'projects/sunlit-descent-196820/jobs/68116593-53e7-4d77-b9b8-fa59fc96cf1a'.

Utilizando o Cloud Trace e o Status do Google Cloud

O Google Cloud fornece ferramentas de diagnóstico que os desenvolvedores de software podem usar para coletar informações sobre o desempenho e o funcionamento de suas aplicações. Especificamente, os desenvolvedores podem usar o Cloud Trace para coletar dados enquanto suas aplicações são executadas.

Visão Geral do Cloud Trace

O Cloud Trace é um sistema de rastreamento distribuído para coletar dados de latência de uma aplicação. Isso ajuda os desenvolvedores a entender onde as aplicações estão gastando seu tempo e a identificar casos em que o desempenho está se degradando.

Do console do Cloud Trace, você pode listar rastreamentos gerados por aplicações que estão rodando em um projeto. Rastreamentos são gerados quando os desenvolvedores chamam especificamente o Cloud Trace de suas aplicações. Além de ver listas de rastreamentos, você pode criar relatórios.

Para o propósito do Exame de Engenharia de Nuvem Associado, lembre-se de que o Cloud Trace é uma aplicação de rastreamento distribuído que ajuda desenvolvedores e engenheiros de DevOps a identificar seções de código que são gargalos de desempenho.

Visualizando o Status do Google Cloud

Além de entender o estado de suas aplicações e serviços, você precisa estar ciente do status dos serviços do Google Cloud. Você pode encontrar esse status no Painel de Status do Google Cloud, que exibe informações sobre o status do serviço: Disponível, Interrupção do Serviço ou Falha do Serviço.

Para visualizar o status dos serviços do Google Cloud, navegue até <https://status.cloud.google.com>. A Figura 18.22 mostra o status geral de áreas geográficas principais.

FIGURE 18.22 Overview status of Google Cloud services

This screenshot shows the 'Service Health' overview for Google Cloud services. At the top, it says 'Check status by product and location. Click the other tabs to check the status for specific regions and multi-regions.' Below this, there's a note about 'Multi-regions' and 'Global' services. A legend indicates: Available (green checkmark), Service information (blue info icon), Service disruption (orange circle with exclamation), and Service outage (red circle with cross). The main table has columns for 'Products' and 'Regions': Americas (regions), Europe (regions), Asia Pacific (regions), and Multi-regions. Products listed include Access Approval, Access Context Manager, Access Transparency, and AI Platform Prediction, all marked as available.

Products	Americas (regions)	Europe (regions)	Asia Pacific (regions)	Multi-regions
Access Approval				
Access Context Manager	Available	Available	Available	Available
Access Transparency				
AI Platform Prediction	Available	Available	Available	Available

Há também abas para ver mais detalhes dentro das principais regiões geográficas. Por exemplo, a Figura 18.23 mostra mais detalhes sobre o status das regiões americanas.

Usando a Calculadora de Preços

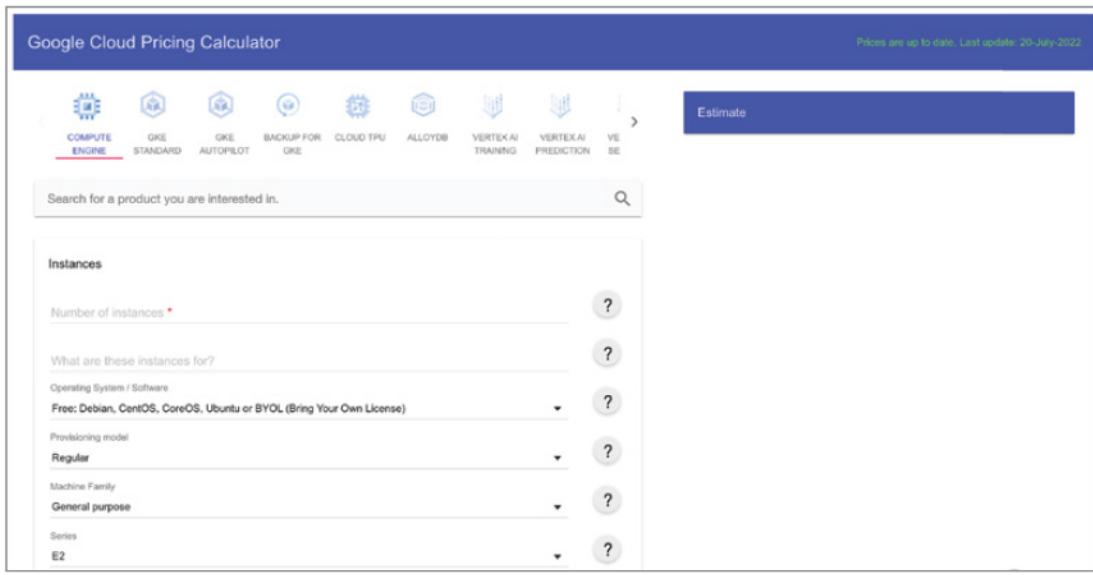
O Google fornece uma Calculadora de Preços para ajudar os usuários do Google Cloud a entender os custos associados aos serviços e configuração de recursos que escolhem usar. Você encontrará a Calculadora de Preços em <https://cloud.google.com/products/calculator> (veja a Figura 18.24).

FIGURE 18.23 More detailed view of American service status

This screenshot shows the 'Americas (regions)' tab of the Google Cloud Service Health page. It displays the status of various services across different US regions: us-central1 (Iowa), us-east1 (South Carolina), us-east4 (Northern Virginia), us-east5 (Columbus), us-south1 (Dallas), and us-west1 (Oregon). All services listed (Access Context Manager, AI Platform Prediction, AI Platform Training, Anthos Service Mesh) are marked as available in all regions.

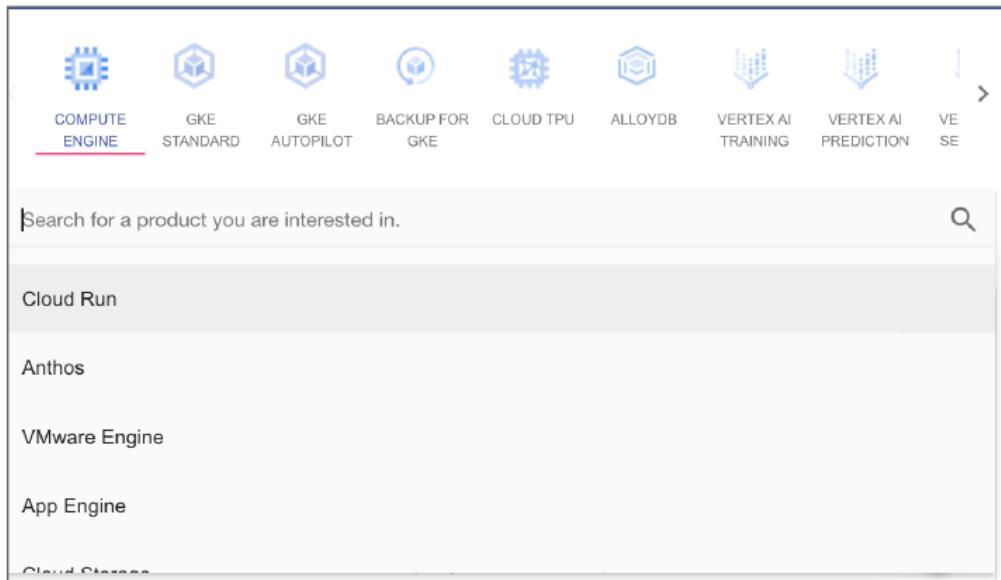
Products	us-central1 Iowa	us-east1 South Carolina	us-east4 Northern Virginia	us-east5 Columbus	us-south1 Dallas	us-west1 Oregon
Access Context Manager	Available	Available	Available	Available		Available
AI Platform Prediction	Available	Available	Available			Available
AI Platform Training	Available	Available	Available			Available
Anthos Service Mesh	Available	Available	Available	Available	Available	Available

FIGURE 18.24 Google Cloud Pricing Calculator



Com a Calculadora de Preços, você pode especificar a configuração de recursos, o tempo que serão usados e, no caso de armazenamento, a quantidade de dados que serão armazenados. Outros parâmetros também podem ser especificados. Eles variarão de acordo com o serviço para o qual você está calculando as taxas. A Figura 18.25 mostra alguns dos serviços disponíveis para usar com a Calculadora de Preços.

FIGURE 18.25 Partial list of services available in the Pricing Calculator



Após selecionar um serviço, você pode especificar uma configuração específica para esse serviço. Por exemplo, ao estimar o preço de uma máquina virtual do Compute Engine, você fornecerá:

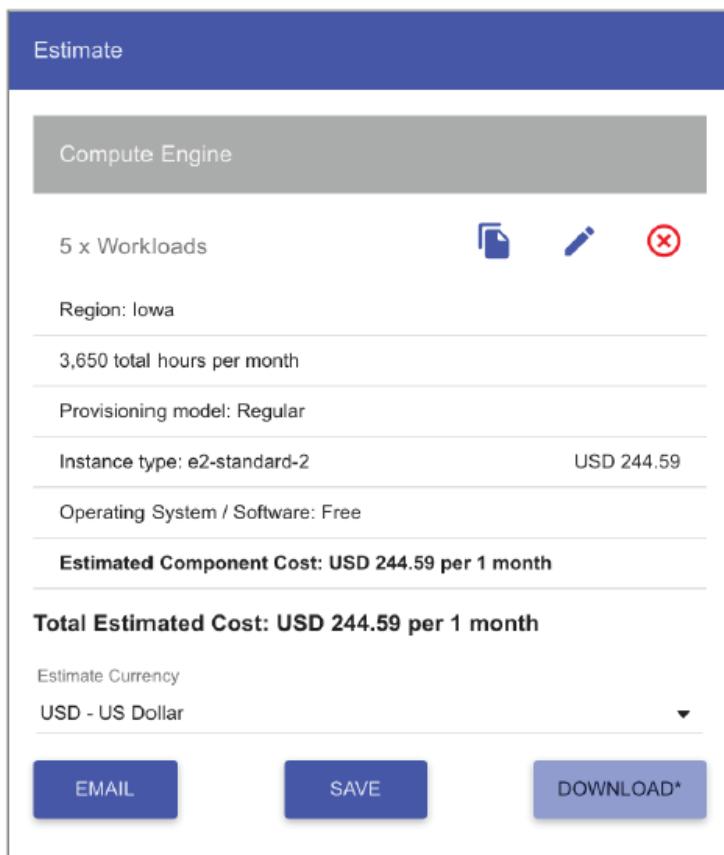
- Número de instâncias
- Tipos de máquina

- Sistema operacional
- Uso médio por dia e por semana
- Discos persistentes
- Balanceamento de carga
- Unidades de processamento tensorial na nuvem (TPUs) (para aplicações de aprendizado de máquina)

Depois de inserir dados nos campos, a Calculadora de Preços gerará uma estimativa, como a mostrada na Figura 18.26.

Recursos diferentes exigirão diferentes parâmetros para uma estimativa. Por exemplo, ao estimar o preço do uso do BigQuery, você precisará especificar tanto parâmetros de armazenamento quanto de consulta. Os parâmetros de armazenamento são para armazenamento ativo e de longo prazo, bem como o volume de dados em inserções e leituras de streaming. Para consultas, você precisará especificar o volume de dados consultados porque o BigQuery cobra com base na quantidade de dados processados ou escaneados para obter resultados de consultas. Atualmente, os primeiros 1 TB de dados processados durante consultas por mês são gratuitos.

FIGURE 18.26 Example price estimate for five e2-standard-2 VMs



Resumo

Como engenheiro de nuvem, você é responsável por monitorar a saúde e o desempenho de aplicações e serviços de nuvem. O Google Cloud fornece várias ferramentas, incluindo serviços de monitoramento, registro e rastreamento.

O Monitoramento na Nuvem permite que você defina alertas em métricas, como a utilização da CPU, para que você possa ser notificado se parte de sua infraestrutura não estiver funcionando conforme o esperado. O Registro na Nuvem coleta, armazena e gerencia entradas de log. Os logs podem ser armazenados em buckets fornecidos pelo Registro na Nuvem ou buckets definidos pelo usuário. Mensagens de log podem ser roteadas para o Cloud Storage, BigQuery ou Cloud Pub/Sub. O Cloud Trace oferece serviços de rastreamento distribuído para identificar partes do código de execução lenta.

Você sempre pode obter o status dos serviços do Google Cloud no Painel de Status do Google Cloud em <https://status.cloud.google.com>.

A Calculadora de Preços é projetada para ajudá-lo a estimar o custo dos serviços no Google Cloud. Está disponível em <https://cloud.google.com/products/calculator>.

Essenciais para o Exame

Entenda a necessidade de monitoramento e o papel das métricas. Métricas fornecem dados sobre o estado das aplicações e infraestrutura. Você cria condições, como a CPU excedendo 80 por cento por 5 minutos, para disparar alertas. Alertas são entregues por canais de notificação. O Google Cloud possui um número substancial de métricas predefinidas, mas você também pode criar métricas personalizadas.

Saiba como coletar, armazenar, filtrar e exibir dados de log usando o Registro na Nuvem. Logs podem vir de praticamente qualquer fonte. O Registro na Nuvem mantém dados de log no bucket Padrão por 30 dias, a menos que uma política de retenção personalizada seja especificada. Se você precisar manter dados de log por mais tempo, precisará exportar os dados para um sink de log. Sinks de log podem ser um bucket do Cloud Storage, um conjunto de dados do BigQuery ou um tópico do Cloud Pub/Sub.

Saiba como filtrar logs. Logs podem conter uma grande quantidade de dados. Use filtros para procurar por texto ou etiquetas, limitar entradas de log por tipo de log e severidade, e restringir o intervalo de tempo para um período de interesse.

As entradas de log são hierárquicas. O Registro na Nuvem mostra um resumo de uma linha para uma entrada de log por padrão, mas você pode explorar os detalhes de uma entrada de log. Use as opções Expandir Tudo e Recolher Tudo para visualizar rapidamente ou ocultar os detalhes completos de uma entrada de log.

Saiba como usar o serviço de rastreamento distribuído Cloud Trace. Desenvolvedores de software incluem código do Cloud Trace em suas aplicações para registrar dados de rastreamento. Dados de rastreamento podem ser visualizados como rastreamentos individuais, ou você pode criar relatórios que incluem parâmetros especificando um subconjunto de rastreamentos que deseja incluir.

Saiba onde o Google Cloud publica o status dos serviços. A página de Status do Google Cloud inclui uma lista de todos os serviços, seu status atual e o status no passado recente. Se houver um incidente em um serviço, você encontrará detalhes adicionais sobre o impacto e a causa raiz do problema.

Saiba como usar a Calculadora de Preços para estimar o custo de recursos e serviços no Google Cloud. A calculadora está disponível em <https://cloud.google.com/products/calculator>. Há uma calculadora separada para cada serviço. Cada serviço tem seu próprio conjunto de parâmetros para estimar custos. A Calculadora de Preços permite estimar o custo de múltiplos serviços e gerar uma estimativa total para todos esses serviços.

Questões de Revisão

1. Qual serviço de Operações na Nuvem é usado para gerar alertas quando a utilização da CPU de uma VM excede 80 por cento?
 - A. Cloud Logging
 - B. Cloud Monitoring
 - C. Cloud Trace
 - D. Cloud Debugger
2. Você acabou de criar uma máquina virtual e gostaria de coletar métricas detalhadas sobre a VM. O que você precisa fazer na VM para que isso aconteça?
 - A. Instalar uma imagem de Cloud Operations.
 - B. Instalar o Agente de Operações na VM.
 - C. Editar a configuração da VM no Cloud Console e selecionar a opção Monitorar com Cloud Monitoring.
 - D. Configurar um canal de notificação.
3. Onde o Cloud Monitoring pode ser usado para monitorar recursos?
 - A. Apenas no Google Cloud
 - B. No Google Cloud e na Amazon Web Services apenas
 - C. No Google Cloud e em data centers locais
 - D. No Google Cloud, Amazon Web Services e data centers locais
4. Você é responsável pela confiabilidade e disponibilidade de vários serviços executados no Kubernetes Engine. Você determinou que precisa monitorar várias métricas para obter informações sobre o estado dos serviços. Você gostaria de ver todas essas métricas exibidas como gráficos de linha, um para cada métrica. Todos os gráficos de linha devem estar disponíveis em uma visualização de página única. O que você usaria para criar tal visualização de página?
 - A. Dashboard do Cloud Monitoring
 - B. Sink do Cloud Logging
 - C. Alerta do Cloud Monitoring
 - D. Conjunto de dados do BigQuery
5. Você criou uma condição de utilização da CPU e quer receber notificações. Quais das seguintes opções estão disponíveis?
 - A. Apenas e-mail
 - B. Apenas PagerDuty

- C. Webhooks e PagerDuty
 - D. E-mail, PagerDuty e Webhooks
6. Quando você cria uma política para notificá-lo de um problema potencial com sua infraestrutura, você pode especificar documentação opcional. Por que se dar ao trabalho de incluir documentação nessa forma?
- A. É salvo no Cloud Storage para uso futuro.
 - B. Pode ajudar você ou um colega a entender o propósito da política.
 - C. Pode conter informações que ajudariam alguém a diagnosticar e corrigir o problema.
 - D. Opções B e C.
7. O que é fadiga de alerta e por que é um problema?
- A. Muitas notificações de alerta são enviadas para eventos que não requerem intervenção humana, e eventualmente os engenheiros de DevOps começam a prestar menos atenção às notificações.
 - B. Muitos alertas colocam carga desnecessária em seus sistemas.
 - C. Poucos alertas deixam os engenheiros de DevOps incertos sobre o estado de suas aplicações e infraestrutura.
 - D. Muitas mensagens de log tornam difícil encontrar mensagens importantes.
8. Por quanto tempo os dados de log são armazenados no bucket Default do Cloud Logging?
- A. 7 dias
 - B. 15 dias
 - C. 30 dias
 - D. 60 dias
9. Você precisa armazenar entradas de log por um período mais longo do que o Cloud Logging as retém no bucket Default. Qual é a melhor opção para preservar os dados de log?
- A. Não há opção; uma vez passado o período de retenção de dados, o Cloud Logging deleta os dados.
 - B. Criar um bucket definido pelo usuário e configurar uma política de retenção.
 - C. Escrever um script em Python para usar a API do Cloud Logging para escrever os dados no Cloud Storage.
 - D. Escrever um script em Python para usar a API do Cloud Logging para escrever os dados no BigQuery.
- Quais das seguintes opções são possíveis para sinks de log?

- A. Apenas bucket do Cloud Storage
 - B. Conjunto de dados do BigQuery e apenas bucket do Cloud Storage
 - C. Apenas tópico do Cloud Pub/Sub
 - D. Bucket do Cloud Storage, conjunto de dados do BigQuery e tópico do Cloud Pub/Sub
10. Qual das seguintes opções pode ser usada para filtrar entradas de log ao visualizar logs no Cloud Logging?
- A. Apenas consulta de log
 - B. Tipo de recurso e gravidade apenas
 - C. Tempo e gravidade apenas
 - D. Consulta de log, tipo de recurso, gravidade e tempo
11. Qual das seguintes não é um nível de log padrão que pode ser usado para filtrar visualizações de log?
- A. Crítico
 - B. Parado
 - C. Aviso
 - D. Informação
12. Você está visualizando entradas de log e encontra uma que parece suspeita. Você não está familiarizado com esse tipo de entrada de log e quer descobrir o que, especificamente, está em um campo chamado metadataRequest. O que você faria?
- A. Expandir o campo metadataRequest na estrutura JSON da mensagem.
 - B. Ver a mensagem no Metric Explorer.
 - C. Escrever um script em Python para reformatar a entrada de log.
 - D. Clicar no link Mostrar Detalhes ao lado da entrada de log.
13. Qual serviço de Operações na Nuvem é melhor para identificar onde existem gargalos em sua aplicação?
- A. Monitoramento
 - B. Logging
 - C. Trace
 - D. Debugger
14. Há um problema de desempenho em um microsserviço. Você revisou as saídas da aplicação, mas não consegue identificar o problema. Qual serviço de Operações na Nuvem você usaria para obter insights sobre o desempenho dos serviços durante a execução?

- A. Monitoramento
 - B. Logging
 - C. Trace
 - D. Debugger
15. Você acredita que pode haver um problema com o BigQuery na zona us-central. Onde você iria para verificar o status do serviço BigQuery para o acesso mais rápido aos detalhes?
- A. Enviar um e-mail para o Suporte do Google Cloud.
 - B. Verificar <https://status.cloud.google.com>.
 - C. Verificar <https://bigquery.status.cloud.google.com>.
 - D. Ligar para o suporte técnico do Google.
16. Você gostaria de estimar o custo dos recursos do Google Cloud que estará usando. Quais serviços exigiriam que você tivesse informações sobre as máquinas virtuais que estará usando?
- A. Compute Engine e BigQuery
 - B. Compute Engine e Kubernetes Engine
 - C. BigQuery e Kubernetes Engine
 - D. BigQuery e Cloud Pub/Sub
17. Você está gerando uma estimativa do custo de usar o BigQuery. Um dos parâmetros é o Preço por Consulta. Você tem que especificar um valor em unidades de TB. Qual é o valor que você está especificando?
- A. A quantidade de dados armazenados no BigQuery
 - B. A quantidade de dados retornados pela consulta
 - C. A quantidade de dados examinados pela consulta
 - D. O número de partições usadas
18. Por que você precisa especificar o sistema operacional a ser usado ao estimar o custo de uma VM?
- A. Todos os sistemas operacionais têm uma taxa fixa.
 - B. Alguns sistemas operacionais incorrem em custo.
 - C. Não é necessário; só é incluído para documentação.
 - D. Para estimar o custo de configurações Traga Sua Própria Licença.
19. Que tipos de mensagens de log são enviadas para o sink de log Requerido?
- A. Apenas mensagens do sistema operacional

- B. Apenas mensagens de atividade de administração
- C. Atividade de administração e eventos do sistema apenas
- D. Atividade de administração, eventos do sistema e transparência de acesso

Capítulo 1: Visão Geral do Google Cloud

1. B. A resposta correta é B. Uma unidade básica para a compra de recursos de computação no Google Cloud é a máquina virtual (VM). A opção A está incorreta; um cache é um sistema de armazenamento de baixa latência. A opção C está incorreta; um bloco é uma unidade de armazenamento em discos persistentes. A opção D está incorreta; uma sub-rede é uma abstração de rede.
2. D. A resposta correta é D. Ao usar clusters gerenciados, o provedor de nuvem monitorará a saúde dos servidores, também conhecidos como nós, no cluster; configurará a rede entre os nós no cluster; e configurará firewall e outros controles de segurança.
3. B. A resposta correta é B. Cloud Run é uma plataforma sem servidor para executar contêineres, e Cloud Functions é um serviço para executar funções de curta duração em resposta a eventos. Kubernetes Engine é um serviço de cluster gerenciado, e tanto o Kubernetes Engine quanto o Compute Engine exigem que você configure servidores. Nem o Compute Engine nem o Kubernetes são opções sem servidor.
4. B. A resposta correta é B. O armazenamento de objetos, como o Cloud Storage, fornece objetos armazenados de forma redundante sem limites na quantidade de dados que você pode armazenar, o que torna a opção B correta. Como a funcionalidade de sistema de arquivos não é necessária, a opção D não é uma boa opção. O armazenamento em bloco poderia ser usado, mas você teria que gerenciar sua própria replicação para garantir alta disponibilidade e isso custaria mais do que o armazenamento de objetos. Caches são armazenamentos transitórios em memória e não são sistemas de armazenamento persistentes de alta disponibilidade.
5. D. A resposta correta é D. Os tamanhos dos blocos em um sistema de armazenamento em bloco podem variar. O tamanho do bloco é estabelecido quando um sistema de arquivos é criado. No Linux, tamanhos de bloco de 4 KB são comumente usados.
6. C. A resposta correta é C. Firewalls no Google Cloud são controles de rede definidos por software que limitam o fluxo de tráfego para dentro e para fora de uma rede ou sub-rede. Roteadores são usados para mover o tráfego para destinos apropriados na rede. Gerenciamento de acesso de identidade é usado para autenticar e autorizar usuários; não é relevante para controles de rede entre sub-redes. Tabelas de endereço IP não são um controle de segurança.
7. C. A opção C está correta porque serviços especializados no Google Cloud, como o AutoML, são sem servidor. O Google gerencia os recursos de computação usados pelos serviços. Não há necessidade de um usuário alocar ou gerenciar servidores.
8. B. A opção B está correta; investir em servidores funciona bem quando uma organização pode prever com precisão o número de servidores e outros equipamentos de que precisará por um período prolongado e pode utilizar esse

equipamento de forma consistente. Startups não são negócios estabelecidos com históricos que podem guiar as necessidades esperadas em três a cinco anos. Não importa se um orçamento é fixo ou variável; investir em servidores deve ser baseado na demanda por capacidade de servidor.

9. B. As características do servidor, como o número de servidores virtuais, a quantidade de memória e a região onde você executa a VM, influenciam o custo, então a opção B está correta. A hora do dia não é um fator, nem o tipo de aplicativo que você executa na VM.
10. D. AutoML é um dos serviços especializados do Google Cloud. Usuários do serviço não precisam configurar nenhuma VM para usar o serviço.
11. B. Os contêineres oferecem a maior flexibilidade para usar os recursos de um cluster de forma eficiente, e plataformas de orquestração reduzem a sobrecarga operacional, o que torna a opção B correta. Não é recomendado executar em uma única VM, porque se o servidor falhar, todos os serviços serão interrompidos. Usar duas VMs com uma somente leitura não é útil. Servidores somente leitura às vezes são usados com bancos de dados, mas não houve menção a bancos de dados na questão. Usar uma VM pequena e atualizar quando não for mais capaz de acompanhar a carga de trabalho oferece um serviço de má qualidade aos usuários e deve ser evitado.
12. D. A resposta correta é D. Todas as operações estão disponíveis para um administrador de sistema após a criação de uma VM.
13. A. A opção A está correta; Cloud Filestore é baseado no Sistema de Arquivos de Rede (NFS), que é um sistema de gerenciamento de arquivos distribuído. As outras opções são sistemas de arquivos suportados pelo Linux.
 14. A. Quando você cria recursos, eles são criados dentro de um VPC. Recursos são adicionados ao VPC e não são acessíveis fora do VPC, a menos que você configure explicitamente para que sejam um subdomínio está relacionado a domínios da web e não à organização de recursos do Google Cloud. Clusters, como clusters Kubernetes, podem estar na sua rede, mas nem todos os recursos necessariamente estão em um cluster.
15. D. A resposta correta é D. Caches usam memória, e isso os torna o tipo de armazenamento mais rápido para leitura de dados. Caches são armazenamentos de dados no backend de sistemas distribuídos, não nos clientes. Um cache não teria efeito na execução de JavaScript do lado do cliente. Caches podem perder dados no cache se a energia for perdida e os dados teriam que ser recarregados. Caches podem ficar desincronizados com o sistema de verdade porque o sistema de verdade poderia ser atualizado, mas o cache pode não ser atualizado. Caches têm tempos de leitura mais rápidos que SSDs e HDDs.
16. B. A opção B está correta; provedores de nuvem têm grande capacidade e podem alojar rapidamente esses recursos para diferentes clientes. Com uma mistura de clientes e cargas de trabalho, eles podem otimizar a alocação de recursos. A opção A está incorreta; provedores de nuvem não tiram recursos de um cliente para dar

a outro, com exceção de instâncias preemptíveis. A opção C está incorreta; provedores de nuvem geralmente oferecem descontos para uso aumentado.

17. C. A opção C está correta. Serviços especializados são monitorados pelo Google para que os usuários não precisem monitorá-los. Serviços especializados fornecem uma funcionalidade de computação específica, mas não requerem que o usuário configure recursos. Eles também fornecem APIs.
18. B. A resposta correta é B. Drives anexados são dispositivos de armazenamento em bloco. Cloud Storage é um serviço de armazenamento de objetos e não se anexa diretamente a uma VM. NoSQL é um tipo de banco de dados, não um sistema de armazenamento. Não existe algo como armazenamento SQL; SQL é uma linguagem de consulta usada em bancos de dados relacionais.
19. C. A resposta correta é C. Bancos de dados requerem armazenamento persistente em dispositivos de bloco. Armazenamento de objetos não fornece armazenamento de bloco de dados ou sistema de arquivos. Armazenamento de dados não é um tipo de sistema de armazenamento. Caches são frequentemente usados com bancos de dados para melhorar o desempenho de leitura, mas são voláteis e não são adequados para armazenar arquivos de dados de forma persistente.
20. B. A resposta correta é B. Todos os três serviços são sem servidor, então o usuário não precisa configurar VMs. Cloud Storage é cobrado com base no tempo e no tamanho dos dados armazenados. Cloud Run e Cloud Functions não são restritos apenas à linguagem Go.

Capítulo 2: Serviços de Computação em Nuvem do Google

1. C. A resposta correta é C. O Cloud Load Balancing distribui cargas de trabalho dentro e entre regiões, oferece verificações de saúde e implementa autoescala. O Cloud DNS fornece serviços de nome de domínio, como traduzir uma URL como www.exemplo.com para um endereço IP. O Cloud Spanner é um banco de dados relacional distribuído, mas não implementa distribuição de carga de trabalho. O Cloud CDN distribui conteúdo através de regiões para reduzir a latência ao entregar conteúdo a usuários em todo o mundo.
2. C. A resposta correta é C. O Cloud Run permite que você execute contêineres em um serviço sem servidor. O Kubernetes Engine é uma plataforma de orquestração para executar contêineres. Ambos fornecem serviços de gerenciamento de contêineres e suportam aplicações com estado. O Cloud Run permite executar contêineres em um serviço gerenciado, mas atualmente não suporta o gerenciamento de estado dentro do contêiner. O ambiente padrão do App Engine executa aplicações em caixas de areia específicas de linguagem e não é um sistema geral de gerenciamento de contêineres. O Cloud Functions é um serviço sem servidor para executar código em resposta a eventos.
3. D. A resposta correta é D. As opções A e B estão corretas. A plataforma API do Apigee fornece serviços de limitação de taxa e roteamento baseados em políticas para ajudar a acomodar picos de tráfego. Também fornece autenticação OAuth 2.0 e SAML. Não fornece controle de versão; o Cloud Source Repositories é o serviço usado para controle de versão.
4. A. A resposta correta é A. O Cloud Armor se baseia nos serviços de balanceamento de carga do Google Cloud para fornecer a capacidade de permitir ou restringir o acesso com base no endereço IP, implantar regras para contrariar ataques de cross-site scripting e fornecer contramedidas para ataques de injeção SQL. O Cloud CDN é um serviço de distribuição de conteúdo, não um serviço de segurança. O gerenciamento de identidade e acesso é um serviço de segurança, mas é para autorização, não para mitigação de ataques de negação de serviço. Redes privadas virtuais são usadas para restringir o acesso à rede aos recursos de uma organização, mas não possuem recursos para mitigar ataques de negação de serviço. Além disso, o Cloud CDN atua como uma primeira linha de defesa no caso de ataques DDoS.
5. A. A resposta correta é A. Este é um bom caso de uso para VMs preemptíveis, pois elas poderiam reduzir o custo de execução da segunda aplicação sem o risco de perder trabalho. Uma vez que as tarefas são excluídas da fila apenas depois de concluídas, se uma VM preemptível for desligada antes de completar a tarefa, outra VM pode realizar a tarefa. Além disso, não há dano em executar uma tarefa mais de uma vez, então se duas VMs fizerem a mesma tarefa, isso não afetará adversamente a saída da aplicação. O DataProc é um cluster gerenciado de Hadoop e Spark e o Spanner é um banco de dados relacional escalável globalmente; nenhum dos dois é apropriado para esta tarefa.

6. B. A resposta correta é B. O Cloud Memorystore é o serviço gerenciado do Google Cloud para cache de dados em memória usando Redis ou memcached. O Cloud SQL é um serviço de banco de dados relacional e pode ser uma boa opção para o banco de dados de back-end. O Cloud Spanner é um banco de dados relacional global e é uma boa opção quando você precisa de um banco de dados relacional escalável globalmente. O Cloud Firestore é um banco de dados de documentos adequado para catálogos de produtos, perfis de usuário e outros dados semi-estruturados.
7. D. A resposta correta é D. Todos os três serviços listados, Compute Engine, Cloud Storage e firewalls de rede, podem ser gerenciados e configurados usando o Cloud SDK.
8. D. A resposta correta é D. O Cloud Functions é um produto sem servidor, então nenhuma configuração é necessária.
9. D. A resposta correta é D. O serviço Cloud Logging é usado para consolidar e gerenciar logs gerados por aplicações e servidores.
10. B. A resposta correta é B. O conjunto de serviços de análise de dados inclui produtos que ajudam com extração, transformação e carregamento (ETL) e trabalham tanto com dados em lote quanto em tempo real. A plataforma API do Apigee é usada para gerenciar APIs e não atende às necessidades descritas. IA e aprendizado de máquina podem ser úteis para analisar dados no armazém de dados, mas os serviços desse conjunto nem sempre são úteis para operações de ETL. O Cloud SDK é usado para controlar serviços, mas por si só não é capaz de realizar as operações necessárias.
11. B. A resposta correta é B. O Bigtable é projetado para aceitar bilhões de linhas de dados. O Spanner é um banco de dados relacional e suporta transações, mas elas não são necessárias. O Cloud SQL MySQL e o Cloud SQL PostgreSQL seriam difíceis de escalar para este nível de desempenho de leitura e escrita.
12. A. A resposta correta é A. O Cloud Firestore é um serviço de banco de dados que pode sincronizar dados entre dispositivos móveis e armazenamento centralizado. O Spanner é um banco de dados relacional global para aplicações de grande escala que requerem suporte a transações em bancos de dados altamente escaláveis. O Cloud CDN é um sistema de armazenamento distribuído para reduzir a latência ao entregar conteúdo estático para usuários de aplicações web. O Cloud SQL poderia ser usado, mas exigiria mais desenvolvimento personalizado para sincronizar dados entre dispositivos móveis e o armazenamento de dados centralizado.
13. B. A resposta correta é B. Uma aplicação intensiva em computação obviamente requer alta CPU, mas o fato de haver muitos cálculos de ponto flutuante indica que uma GPU deve ser usada. Você pode considerar executar isso em um cluster, mas o trabalho não é facilmente distribuído em vários servidores, então você precisará ter um servidor único capaz de lidar com a carga. O acesso imediato a grandes quantidades de dados indica que uma máquina de alta memória deve ser recomendada.

14. B. A resposta correta é B. Identidades são abstrações de usuários. Elas também podem representar características de processos que executam em nome de um usuário humano ou de uma VM no Google Cloud; esses são conhecidos como contas de serviço. Identidades não estão relacionadas a IDs de VM. Funções são coleções de privilégios que podem ser concedidos a identidades. A opção D é sinônima da opção C.
15. C. A resposta correta é C. Os serviços de Linguagem Natural fornecem funcionalidades para analisar texto. O Vertex AI é uma plataforma unificada para construir modelos de aprendizado de máquina, mas como o cliente não é um especialista em aprendizado de máquina, um serviço especializado como o de Linguagem Natural é uma opção melhor. O Recommendation AI é usado para fazer recomendações de produtos para clientes. Texto para Fala é um serviço para converter texto de linguagem natural em fala soando humana.
16. B. A resposta correta é B. Ambas as opções B e D atenderiam à necessidade de executar Spark, o que daria aos cientistas de dados acesso à biblioteca de máquina que eles precisam. No entanto, a opção D exige que eles gerenciem e monitorem o cluster de servidores, o que exigiria mais trabalho de DevOps e administração do que se usassem o serviço Dataproc. A opção C, BigQuery, é um banco de dados escalável, não uma plataforma para executar Spark. Cloud Spark é um produto fictício e não existe no Google Cloud.
17. B. A resposta correta é B. O Spanner suporta ANSI SQL 2011 e transações globais. O Cloud SQL suporta SQL padrão, mas não tem transação global. O Firestore e o Bigtable são bancos de dados NoSQL.
18. A. O Dataproc é projetado para executar workflows tanto em modos batch quanto em streaming, o que torna a opção A correta. O BigQuery é um serviço de armazém de dados. O Firestore é um banco de dados de documentos. O AutoML é um serviço de aprendizado de máquina.
19. C. A resposta correta é C. O ambiente padrão do App Engine fornece uma caixa de areia Python sem servidor que escala automaticamente. O ambiente flexível do App Engine executa contêineres e requer mais configuração. O Cloud Engine e o Kubernetes Engine ambos requerem gerenciamento e monitoramento significativos.]
20. D. A resposta correta é D. O relatório de erros consolida informações de falhas. O Cloud Monitoring coleta métricas sobre desempenho de aplicativos e servidores. O Logging é um serviço de gerenciamento de logs. O Cloud Dataproc não é uma ferramenta de observabilidade, mas é um serviço gerenciado de Hadoop e Spark.

Capítulo 3: Projetos, Contas de Serviço e Faturamento

1. A opção A, a resposta correta, separa as duas principais aplicações em suas próprias pastas e ainda permite separar seguros privados de pagadores governamentais usando pastas para cada um. Isso satisfaz a necessidade regulatória de manter o software do pagador governamental isolado de outros softwares. A opção B não inclui uma organização, que é a raiz da hierarquia de recursos. A opção C não é flexível em relação às diferenças nas restrições em diferentes aplicações. A opção D é falsa porque a opção A atende aos requisitos.
2. C. Hierarquias de recursos têm uma única organização na raiz, o que torna a opção C correta. Abaixo disso, existem pastas que podem conter outras pastas ou projetos. Pastas podem conter múltiplas pastas e múltiplos projetos.
3. B. Contas de serviço são projetadas para dar permissões a aplicações ou VMs para realizar tarefas. Contas de faturamento são para associar encargos com um método de pagamento. Pastas são parte da hierarquia de recursos e não têm nada a ver com habilitar uma aplicação para realizar uma tarefa. Contas de mensagens são uma opção fictícia.
4. A. Políticas herdadas podem ser sobrepostas definindo uma política em nível de pasta ou projeto. Contas de serviço e contas de faturamento não fazem parte da hierarquia de recursos e não estão envolvidas na sobreposição de políticas.
5. E. Todos os tipos listados de restrições são suportados em políticas.
6. B. A opção B é a resposta correta porque Publisher não é um papel primitivo. Owner (Proprietário), Editor (Editor) e Viewer (Visualizador) são os três papéis básicos no Google Cloud.
7. D. Papéis básicos incluem apenas as permissões de Proprietário, Editor e Visualizador. Papéis predefinidos são projetados para produtos e serviços do Google Cloud, como App Engine e BigQuery. Para uma aplicação personalizada, você pode criar conjuntos de privilégios que dão ao usuário com esse papel tanto permissão quanto necessário, mas não mais.
8. D. Usuários devem ter apenas os privilégios necessários para realizar suas funções. Este é o princípio do menor privilégio. Rotação de funções é outro princípio de segurança relacionado a ter diferentes pessoas realizando uma tarefa em diferentes tempos. Defesa em profundidade é a prática de usar múltiplos controles de segurança para proteger o mesmo ativo. A opção B não é um princípio de segurança real.
9. A. Uma hierarquia de recursos tem apenas uma organização, o que torna a opção A correta. No entanto, você pode criar múltiplas pastas e projetos dentro de uma hierarquia de recursos.
10. B. Na opção B, a resposta correta, a conta de faturamento é usada para especificar informações de pagamento e deve ser usada para configurar pagamentos automáticos. Contas de serviço são usadas para conceder privilégios a uma VM e

não estão relacionadas a faturamento e pagamentos. Contas de recursos e contas de crédito não existem.

11. C. O Google Cloud oferece um nível de serviço gratuito para muitos produtos, o que torna a opção C a resposta correta. Você pode usar esses serviços sem ter que configurar uma conta de faturamento. O Google cobra por produtos sem servidor, como Cloud Functions e App Engine, quando os clientes excedem a quantidade permitida sob o nível gratuito.
12. C. A resposta correta é C. Orçamento e Alertas permitem especificar um orçamento. Quando porcentagens especificadas desse orçamento são gastas, alertas podem ser gerados. O Cloud Monitoring é um serviço de observabilidade para desempenho de aplicativos e infraestrutura, não para faturamento. O Cloud Logging é um serviço de observabilidade para coletar informações sobre eventos em serviços e infraestrutura. Restrições de Política são um mecanismo para restringir como os recursos podem ser usados.
13. D. Grandes empresas devem usar faturamento por fatura quando incorrem em grandes encargos, o que torna a opção D a resposta correta. Uma conta de autoatendimento é apropriada apenas para quantias que estão dentro dos limites de crédito dos cartões de crédito. Uma vez que as subdivisões são gerenciadas independentemente e têm seus próprios orçamentos, cada uma deve ter suas próprias contas de faturamento.
14. A. Quando um usuário recebe a permissão `iam.serviceAccountUser` no nível do projeto, esse usuário pode gerenciar todas as contas de serviço no projeto, então a opção A está correta. Se uma nova conta de serviço for criada, eles automaticamente terão privilégio para gerenciar essa conta de serviço. Você poderia conceder `iam.serviceAccountUser` ao administrador no nível da conta de serviço, mas isso exigiria definir o papel para todas as contas de serviço. Se uma nova conta de serviço for criada, o administrador da aplicação teria que conceder `iam.serviceAccountUser` ao outro administrador na nova conta de serviço. `iam.serviceProjectAccountUser` é um papel fictício.
15. C. Quando uma conta de serviço é criada, o Google gera chaves de criptografia para autenticação, tornando a opção C correta. Nomes de usuário e senhas não são uma opção para contas de serviço. Autenticação de dois fatores é uma prática de autenticação que requer duas formas de autenticação, como um par de nome de usuário e senha e um código de um dispositivo de autenticação. Biometria não pode ser usada por serviços e não é uma opção.
16. B. Contas de serviço são recursos que são gerenciados por administradores, mas também funcionam como identidades que podem ser atribuídas a papéis, o que torna a opção B a resposta correta. Contas de faturamento não estão relacionadas a identidades. Projetos não são identidades; eles não podem assumir papéis. Papéis são recursos, mas não identidades. Eles podem assumir privilégios, mas esses privilégios são usados apenas quando estão anexados a uma identidade.
17. B. Papéis predefinidos são definidos para um produto específico, como Cloud Run ou Compute Engine, então a opção B é a resposta correta. Eles agrupam

privilegios frequentemente necessários juntos ao gerenciar ou usar um serviço. Papéis básicos são blocos de construção para outros papéis. Papéis personalizados são criados por usuários para atender às suas necessidades particulares; o papel Aplicação é um papel fictício.

18. B. Por padrão, todos os usuários em uma organização podem criar projetos, o que torna a opção B correta. O papel resourcemanager.projects.create permite aos usuários criar projetos. A conta de faturamento não está associada à criação de projetos.
19. D. O número máximo de organizações é determinado por conta pelo Google, então a opção D é a resposta correta. Se você precisar de organizações adicionais, pode entrar em contato com o Google e pedir um aumento no seu limite.
20. B. Usuários com o papel de Administrador da Organização não são necessariamente responsáveis por determinar quais permissões devem ser atribuídas aos usuários. Isso é determinado com base no papel da pessoa na organização e nas políticas de segurança estabelecidas dentro da organização, o que torna a opção B correta.

Capítulo 4: Introdução à Computação no Google Cloud

1. B. O ambiente padrão do App Engine pode executar aplicações Python, que podem fazer autoescala para zero instâncias quando não há carga, minimizando assim os custos. O Compute Engine e o ambiente flexível do App Engine requerem mais gerenciamento de configuração do que o ambiente padrão do App Engine. O Kubernetes Engine é usado quando um cluster de servidores é necessário para suportar grandes ou múltiplas aplicações usando os mesmos recursos de computação.
2. A. Servidores de banco de dados requerem alta disponibilidade para responder a consultas de usuários ou aplicações. Máquinas preemptíveis certamente serão desligadas em no máximo 24 horas, a menos que sejam VMs spot. Um trabalho de processamento em lote sem requisitos de tempo fixos poderia usar máquinas preemptíveis desde que a VM seja reiniciada. Clusters de computação de alto desempenho podem usar máquinas preemptíveis porque o trabalho em uma máquina preemptível pode ser automaticamente reagendado para outro nó no cluster quando um servidor é preemptido. A opção D está incorreta porque há uma resposta correta no conjunto de opções.
3. A. VMs são criadas em projetos, que fazem parte da hierarquia de recursos. Elas também estão localizadas em regiões geográficas e data centers, então uma zona é especificada também. Nomes de usuário e papéis de admin não são especificados durante a criação. A conta de faturamento está vinculada a um projeto e, portanto, não precisa ser especificada quando a VM é criada. Buckets de armazenamento em nuvem são criados independentemente de VMs. Nem todas as VMs utilizarão buckets de armazenamento.
4. C. O Compute Engine pode executar contêineres Docker se você instalar o Docker na VM. O Kubernetes e o ambiente flexível do App Engine suportam contêineres Docker. O ambiente padrão do App Engine fornece ambientes de execução específicos de linguagem e não permite que os clientes especifiquem imagens Docker personalizadas para uso.
5. B. O nome do arquivo usado para construir e configurar um contêiner Docker é Dockerfile.
6. D. Anthos é um serviço gerenciado para administrar clusters Kubernetes no Google Cloud, outros clouds e localmente. O App Engine Flexível e o Cloud Functions não são gerenciados pelo Anthos.
7. B. Kubernetes fornece balanceamento de carga, escalabilidade e atualização automática de software. Ele não fornece varredura de vulnerabilidade. O serviço de Scanner de Segurança da Web do Google Cloud e o serviço de Análise de Contêineres podem detectar vulnerabilidades, mas eles são separados do Kubernetes Engine.
8. D. O cenário descrito é adequado para o Kubernetes. Cada um dos grupos de serviços pode ser estruturado em pods e implantado usando o deployment do

Kubernetes. O Kubernetes Engine gerencia a saúde do nó, balanceamento de carga e escalabilidade. O App Engine Edição Padrão possui caixas de areia específicas de linguagem e não é adequado para este caso de uso. O Cloud Functions é projetado para processamento de eventos de curta duração e não é o tipo de processamento contínuo necessário neste cenário. O Compute Engine poderia atender aos requisitos deste caso de uso, mas exigiria mais esforço por parte dos administradores de aplicativos e profissionais de DevOps para configurar平衡adores de carga, monitorar a saúde e gerenciar implantações de software.

9. B. Este é um caso de uso ideal para o Cloud Functions. A função na nuvem é acionada por um evento de upload de arquivo. A função na nuvem chama o serviço de processamento de imagem. Com essa configuração, os dois serviços são independentes. Nenhum servidor adicional é necessário. A opção A viola o requisito de manter os serviços independentes. As opções C e D incorrem em mais sobrecarga de gerenciamento e provavelmente custarão mais para operar do que a opção B.
10. D. Cada invocação de uma função na nuvem é executada em um ambiente de execução seguro e isolado. Não há necessidade de verificar se outras invocações estão em execução. Com o serviço Cloud Functions, não há como um desenvolvedor controlar a execução do código no nível do processo ou thread.
11. A. Você criaria uma imagem personalizada depois de instalar o código personalizado, neste caso, a biblioteca de criptografia. Uma imagem pública não contém código personalizado, mas pode ser usada como base à qual você adiciona código personalizado. Tanto o CentOS quanto o Ubuntu são distribuições Linux. Você poderia usar qualquer um deles como a imagem base à qual você adiciona código personalizado, mas por si só, eles não têm código personalizado.
12. B. Projetos são o nível mais baixo da hierarquia de recursos. A organização está no topo da hierarquia, e as pastas estão entre a organização e os projetos. Instâncias de VM não fazem parte da hierarquia de recursos.
13. D. Todas as regiões do Google têm o mesmo nível de acordo de nível de serviço, portanto, a confiabilidade é a mesma. Os custos podem variar entre as regiões. Regulamentações podem exigir que os dados permaneçam dentro de uma área geográfica, como a União Europeia. A latência é uma consideração quando você deseja uma região que esteja próxima aos usuários finais ou aos dados que você precisará já estarem armazenados em uma região específica.
14. B. O papel de Admin do Compute Engine dá aos usuários controle total sobre as instâncias. As opções A e C são papéis fictícios. O Admin de Segurança do Compute Engine dá aos usuários privilégios para criar, modificar e deletar certificados SSL e regras de firewall.
15. D. VMs preemptíveis serão encerradas após 24 horas, com exceção de VMs spot. O Google não garante que VMs preemptíveis estarão disponíveis. Uma vez que uma instância é iniciada como uma máquina preemptível, ela não pode migrar para uma VM regular. No entanto, você poderia salvar um snapshot e usar isso para criar uma nova instância regular.

16. B. A aplicação mantém estado e, portanto, não pode ser executada no Cloud Run. O Cloud Run é um serviço gerenciado para executar aplicações em contêineres, incluindo contêineres baseados em Docker. Contêineres podem executar aplicações escritas em uma variedade de linguagens.
17. C. A linguagem de programação C não é suportada no ambiente padrão do App Engine. Se você precisar executar uma aplicação C, ela pode ser compilada e executada em um contêiner no ambiente flexível do App Engine.
18. B. O Anthos Service Mesh é um serviço gerenciado que permite serviços consistentes de segurança e monitoramento em clusters Kubernetes. O Cloud Functions é usado para processamento de eventos. O App Engine Standard e o App Engine Flexível são serviços para executar aplicações contêinerizadas.
19. B. O vTPM verifica a integridade de boot de instâncias do Compute Engine e é usado para prevenir rootkits e outro software malicioso de comprometer o sistema operacional. Chaves de criptografia fornecidas pelo cliente são usadas para criptografar dados em repouso. A tenância exclusiva limita quais instâncias podem ser executadas em um servidor, mas não valida a integridade do boot. O gerenciamento de identidade e acesso é usado para atribuir papéis e permissões para controlar o acesso a recursos no Google Cloud.
20. A. O Cloud Functions é mais adequado para processamento baseado em eventos, como um arquivo sendo carregado para o Cloud Storage ou um evento sendo escrito em uma fila Pub/Sub. Trabalhos de longa duração, como carregar dados em um armazém de dados, são mais adequados para o Compute Engine ou App Engine.

Capítulo 5: Computação com Máquinas Virtuais do Compute Engine

1. C. Você deve verificar o projeto selecionado, pois todas as operações que você realizar serão aplicadas aos recursos no projeto selecionado, tornando a opção C a resposta correta. Você não precisa abrir o Cloud Shell, a menos que queira trabalhar com a linha de comando, e, se o fizer, deve verificar primeiro se o projeto está corretamente selecionado. Fazer login em uma VM usando SSH é uma das tarefas que requer que você esteja trabalhando com o projeto correto para ver as VMs associadas a esse projeto, portanto, fazer login via SSH não deve acontecer antes de verificar o projeto. A lista de VMs na janela de Instância de VM é uma lista de VMs no projeto atual. Você deve verificar qual projeto está usando para garantir que está visualizando o conjunto de VMs que pensa estar usando.
2. A. Você precisará configurar o faturamento se ele ainda não estiver habilitado quando começar a usar o console, então a opção A é a resposta correta. Você pode criar um projeto, mas só poderá fazer isso se o faturamento estiver habilitado. Você não precisa criar um bucket de armazenamento para trabalhar com o console. Especificar uma zona padrão não é uma tarefa única; você pode mudar de zonas ao longo da vida do seu projeto.
3. B. O nome da VM, a região e a zona, e o tipo de máquina podem todos ser especificados no console junto com outros parâmetros, então a opção B está correta. A opção A está faltando parâmetros necessários. Um bloco CIDR é uma faixa de endereços IP associada a uma sub-rede e não é necessário para criar uma VM. Um endereço IP é atribuído automaticamente, portanto não é necessário.
4. B. Zonas diferentes podem ter tipos de máquina diferentes disponíveis, então você precisará especificar uma região primeiro e depois uma zona para determinar o conjunto de tipos de máquina disponíveis. Se o tipo de máquina não aparecer na lista, não está disponível nessa zona. Isso torna a opção B a resposta correta. As opções A e C estão incorretas. Sub-redes e endereços IP não estão relacionados aos tipos de máquina disponíveis. A menos que você esteja especificando um tipo de máquina personalizado, você não especifica a quantidade de memória; isso é definido pelo tipo de máquina, então a opção D está incorreta.
5. C. Rótulos e descrições ajudam você a rastrear seus próprios atributos de recursos. À medida que o número de servidores cresce, pode se tornar difícil rastrear quais VMs são usadas para quais aplicações e serviços, então a opção C é a resposta correta. Rótulos e uma descrição geral ajudarão os administradores a rastrear o número de VMs e seus custos relacionados. As opções A e B são usadas para segurança e armazenamento, mas não ajudam na gestão de múltiplas VMs. A opção D está apenas parcialmente correta. Descrições são úteis, mas os rótulos também.
6. A. A seção Política de Disponibilidade dentro da aba Gerenciamento é onde você define a opção preemptível, então a opção A está correta. Identidade e Acesso a API é usada para controlar o acesso da VM às APIs do Google Cloud e qual conta de serviço é usada com a VM. Tenência Única é usada se você precisa executar

susas VMs em servidores físicos que executam apenas suas VMs. Rede é usada para definir tags de rede e alterar a interface de rede.

7. B. VM Blindada é um conjunto avançado de controles de segurança que inclui Monitoramento de Integridade, uma verificação para garantir que as imagens de boot não tenham sido adulteradas, o que torna a opção B a resposta certa. Firewalls são usados para controlar a entrada e saída de tráfego de rede para um servidor ou sub-rede. Chaves SSH em todo o projeto são usadas para autenticar usuários em servidores dentro de um projeto. Controle de integridade do disco de boot é um recurso fictício.
8. C. Tamanho do bloco não é uma opção em Discos Adicionais, então a opção C está correta. Gerenciamento de chave de criptografia, tipo de disco e a opção de especificar uma imagem fonte são todas opções disponíveis.
9. B. Usar scripts controlados por versão é a melhor abordagem das quatro opções. Scripts podem ser documentados com razões para as mudanças e podem ser executados repetidamente em diferentes máquinas para implementar a mesma mudança. Isso reduz a chance de erro ao entrar manualmente um comando. A opção A não ajuda a melhorar a documentação de por que as mudanças foram feitas. A opção C poderia ajudar a melhorar a documentação, mas scripts executáveis são reflexos precisos e exatos do que foi executado. Notas podem perder detalhes. A opção D não é aconselhável. Você poderia se tornar um gargalo para fazer mudanças, mudanças não podem ser feitas quando você está indisponível, e sua memória pode não ser uma forma confiável de rastrear todas as mudanças de configuração.
10. A. gcloud compute instances é o começo de comandos para administrar recursos do Compute Engine, tornando a opção A a resposta certa. A opção B, gcloud instances, está faltando a palavra-chave compute que indica que estamos trabalhando com o Compute Engine. A opção C trocou a ordem de compute e instances. A opção D é falsa porque a opção A é a resposta correta.
11. B. A opção B segue o padrão do comando glcoud, que é hierárquico e começa com o nome do serviço glcoud, neste caso compute para o Compute Engine, seguido pelo próximo nível abaixo, que neste caso é instances. Finalmente, há a ação ou verbo, neste caso list. A opção A está faltando o termo instances para indicar que você está trabalhando com instâncias de VM. A opção C está faltando a palavra-chave compute para indicar que você está trabalhando com o Compute Engine. A opção D está faltando a palavra-chave compute instance e trocou a ordem de instances e list.
12. B. O formato correto é usar o parâmetro --labels e especificar a chave seguida de um sinal de igual seguido pelo valor na opção B. As opções A e C têm o caractere errado separando a chave e o valor. A opção D está incorreta porque é possível especificar rótulos na linha de comando.
13. C. As duas operações que você pode especificar ao usar a configuração de disco de boot são adicionar um novo disco e anexar um disco existente, então a opção

C está correta. Reformatar um disco existente não é uma opção, então as opções A, B e D não podem ser a resposta correta.

14. B. 10 GB de dados são pequenos o suficiente para serem armazenados em um único disco. Criando uma imagem de um disco com os dados armazenados nela, você pode especificar essa imagem fonte ao criar uma VM. A opção A exigiria que o cientista de dados copiasse os dados do Cloud Storage para um disco na VM. A opção C exigiria similarmente a cópia dos dados. A opção D carregaria dados em um banco de dados, não em um sistema de arquivos conforme especificado nos requisitos.
15. B. Na aba de Rede do formulário da VM, você pode adicionar outra interface de rede, então a opção B está correta. O GCP define o endereço IP, então a opção A está incorreta. Não há opção para especificar um roteador ou alterar regras de firewall na aba de Rede, então as opções C e D estão incorretas.
16. A. A opção correta é boot-disk-type, que é a opção A. As outras três opções não são parâmetros do comando gcloud compute instances.
17. A. A opção A é o comando correto. É a única opção que inclui um tipo de máquina correto e especifica adequadamente o nome da instância. A opção B usa o parâmetro --cpus, que não existe. A opção C usa o parâmetro instance-name, que não existe. O nome da instância é passado como um argumento e não precisa de um nome de parâmetro. A opção D está incorreta porque o tipo de máquina n1-4-cpu não é um tipo de máquina válido.
18. C. A opção C é o comando correto, que é gcloud compute instances, para indicar que você está trabalhando com VMs, seguido pelo comando stop e o nome da VM. A opção A está incorreta porque halt não é uma opção. A opção B está incorreta porque terminate não é um parâmetro. A opção D está faltando a palavra instances, que indica que você está trabalhando com VMs.
19. B. SSH é um serviço para conectar-se a um servidor remoto e fazer login em uma janela de terminal. Uma vez logado, você teria acesso a uma linha de comando, então a opção B é a resposta certa. FTP é um protocolo de transferência de arquivos e não permite que você faça login e realize tarefas de administração do sistema. RDP é um protocolo usado para acessar remotamente servidores Windows, não Ubuntu, que é uma distribuição Linux. ipconfig é uma utilidade de linha de comando para configurar pilhas IP em um dispositivo e não permite que você faça login em um servidor remoto.
20. A. Todas as declarações na opção A são verdadeiras e relevantes para faturamento e custos. A opção B está correta que VMs são cobradas em incrementos de 1 segundo, mas apenas VMs preemptíveis são desligadas dentro de 24 horas após o início. A opção C está incorreta porque descontos não são limitados a algumas regiões. A opção D está incorreta porque VMs não são cobradas por um mínimo de 1 hora.

Capítulo 6: Gerenciando Máquinas Virtuais

1. A. A página do Compute Engine é onde você tem a opção de criar uma única instância de VM, então a opção A é a resposta correta. O App Engine é usado para contêineres e execução de aplicações em ambientes de tempo de execução específicos da linguagem. O Kubernetes Engine é usado para criar e gerenciar clusters Kubernetes. O Cloud Functions é onde você criaria uma função para executar no ambiente de função de nuvem sem servidor do Google.
2. B. Instâncias podem ser paradas, e quando estão, então você não pode se conectar a elas via SSH, o que torna a opção B a resposta correta. Iniciar a instância habilitará o acesso SSH. A opção A não está correta porque você pode fazer login em máquinas preemptíveis. A opção C está incorreta porque não existe uma opção Sem SSH. A opção D está incorreta porque a opção SSH pode ser desabilitada.
3. B. O comando Reset pode ser usado para reiniciar uma VM; portanto, a opção B está correta. As propriedades da VM não mudarão, mas os dados na memória serão perdidos. Não há opções de Reboot, Restart, Shutdown ou Startup no console.
4. C. Rótulos, proteção contra exclusão e status estão todos disponíveis para filtragem, então a opção C é a resposta correta. Você também pode filtrar por IP interno, IP externo, zona, rede, proteção contra exclusão e membro de um grupo de instâncias gerenciadas ou não gerenciadas.
5. A. Para funcionar corretamente, o sistema operacional deve ter as bibliotecas de GPU instaladas, então a opção A está correta. O sistema operacional não precisa ser baseado em Ubuntu, e não há necessidade de ter pelo menos oito CPUs em uma instância antes de poder anexar e usar uma GPU. O espaço disponível em disco não determina se uma GPU é usada ou não.
6. A. Se você adicionar uma GPU a uma VM, deve ter CPUs e GPUs compatíveis. A instância não precisa ser preemptível e pode ter discos não bootáveis anexados. A instância não é obrigada a executar Ubuntu 18.02 ou posterior.
7. B. Quando você cria um snapshot pela primeira vez, o Google Cloud fará uma cópia completa dos dados no disco persistente. Na próxima vez que você criar um snapshot daquele disco, o Google Cloud copiará apenas os dados que mudaram desde o último snapshot. A opção A está incorreta; o Google Cloud não armazena uma cópia completa para o segundo snapshot. A opção C está incorreta; o primeiro snapshot não é deletado automaticamente. A opção D está incorreta; snapshots subsequentes não incorrem em sobrecarga de 10%.
8. D. Para trabalhar com snapshots, um usuário deve ser atribuído ao papel de Administrador de Armazenamento do Compute, o que torna a opção D a resposta correta. As outras opções são papéis fictícios.
9. C. Imagens podem ser criadas a partir de discos, snapshots, arquivos de armazenamento na nuvem, um disco virtual ou outra imagem, então a opção C é a resposta correta. Arquivos de exportação de banco de dados não são fontes para imagens.

10. B. Depreciado marca a imagem como não mais suportada e permite que você especifique uma imagem de substituição para usar daí para frente, tornando a opção B a resposta correta. Imagens depreciadas estão disponíveis para uso, mas podem não ser corrigidas para falhas de segurança ou ter outras atualizações. As outras opções são recursos fictícios de imagens.
11. C. O comando base para trabalhar com instâncias é gcloud compute instances, o que torna a opção C a resposta correta. O comando list é usado para mostrar detalhes de todas as instâncias. Por padrão, a saída está em forma legível por humanos, não em json. Usar a opção --format json força a saída a estar no formato JSON. --output não é uma opção válida.
12. B. --async faz com que informações sobre o processo de inicialização sejam exibidas; portanto, a opção B está correta. --verbose é um parâmetro análogo em muitos comandos Linux. --describe fornece detalhes sobre uma instância, mas não necessariamente sobre o processo de inicialização. --details não é um parâmetro válido.
13. C. O comando para deletar uma instância é gcloud compute instances delete seguido pelo nome da instância, então a opção C está correta. A opção A está incorreta porque não há parâmetro instance. A opção B está incorreta porque esse comando para, mas não deleta a instância. A opção D está faltando instances no comando, que é necessário para indicar que tipo de entidade está sendo deletada.
14. A. gcloud compute instances é o comando base seguido por delete, o nome da instância, e --keep-disks=boot, então a opção A está correta. Não existe parâmetro --save-disk. A opção C está errada porque filesystem não é um valor válido para o parâmetro keep-disk. A opção D está faltando a opção instances, que é necessária no comando.
15. B. A resposta correta é a opção B, que é usar o comando describe. A opção A mostrará alguns campos, mas não todos. As opções C e D estão incorretas porque não existe parâmetro detailed.
16. B. Grupos de instâncias são conjuntos de VMs que podem ser configurados para escalar e são usados com平衡adores de carga, o que contribui para melhorar a disponibilidade, então a opção B está correta. Instâncias preemptíveis não são altamente disponíveis porque podem ser desligadas a qualquer momento pelo Google Cloud. O Cloud Storage não é um componente do Compute Engine. GPUs podem ajudar a melhorar o throughput para operações intensivas em matemática, mas não contribuem para alta disponibilidade.
17. B. Um modelo de instância é usado para especificar como o grupo de instâncias deve ser criado, o que torna a opção B a resposta correta. A opção A está incorreta porque instâncias são criadas automaticamente quando um grupo de instâncias é criado. Imagens de disco de boot e snapshots não precisam ser criados antes de criar um grupo de instâncias.
18. B. O comando para deletar um grupo de instâncias é gcloud compute instance-template delete, então a opção B está correta. A opção A inclui incorretamente o

termo instances. A opção C está em ordem incorreta. A opção D está errada porque instance-template está na posição errada e está no plural na opção.

19. C. Você pode configurar uma política de escalonamento automático para acionar a adição ou remoção de instâncias com base na utilização da CPU, métrica de monitoramento, capacidade de balanceamento de carga ou cargas de trabalho baseadas em fila. Disco, latência de rede e memória podem acionar o escalonamento se métricas de monitoramento nesses recursos forem configuradas. Então, a opção C está correta.
20. B. Grupos de instâncias não gerenciados estão disponíveis para casos de uso limitados como este. Grupos de instâncias não gerenciados não são recomendados em geral. Grupos de instâncias gerenciados são a maneira recomendada de usar grupos de instâncias, mas as duas configurações diferentes impedem seu uso. Instâncias preemptíveis e GPUs não são relevantes para este cenário.

Capítulo 7: Computação com Kubernetes

1. C. O Kubernetes cria grupos de instâncias como parte do processo de criação de um cluster, o que torna a opção C a resposta correta. O Cloud Monitoring e o Cloud Logging, não grupos de instâncias, são usados para monitorar a saúde dos nós e criar alertas e notificações. O Kubernetes cria pods e deployments; eles não são fornecidos por grupos de instâncias.
2. A. Um cluster Kubernetes tem um único plano de controle e um ou mais nós para executar cargas de trabalho, então a opção A é a resposta correta. Não há nó de monitoramento no Kubernetes, mas ele gera métricas que podem ser enviadas para o Cloud Monitoring. O Kubernetes não exige instâncias com pelo menos seis vCPUs.
3. C. Pods são instâncias únicas de uma aplicação em execução em um cluster, então a opção C está correta. Pods executam contêineres mas não são simplesmente conjuntos de contêineres. O código da aplicação é executado em contêineres que são implantados em pods. Pods não são controladores, então eles não podem gerenciar comunicação com clientes e serviços do Kubernetes.
4. B. Serviços são componentes do Kubernetes que fornecem pontos finais da API que permitem que as aplicações descubram pods executando uma aplicação específica, tornando a opção B correta. As opções A e C, se pudessem ser codificadas usando a API projetada para gerenciar clusters, exigiriam mais código do que trabalhar com serviços e estão sujeitas a alterações em um conjunto maior de funções da API. A opção D não é uma opção real.
5. A. Uma configuração de deployment especifica quantos nós criar em um ReplicaSet. O Cloud Operations Suite é um serviço de monitoramento e log que monitora mas não controla clusters Kubernetes. O Container Runtime é um componente do Kubernetes responsável por executar contêineres. Jobs é uma abstração de cargas de trabalho e não está vinculada ao número de pods executando em um cluster.
6. B. Clusters regionais estão disponíveis no Kubernetes Engine e são usados para fornecer resiliência a uma aplicação, então a opção B está correta. A opção A refere-se a grupos de instâncias que são um recurso do Compute Engine, não diretamente do Kubernetes Engine. A opção C está incorreta; deployments regionais é um termo fictício. O balanceamento de carga distribui a carga e faz parte do Kubernetes por padrão. Se a carga não for distribuída entre zonas ou regiões, isso não ajuda a adicionar resiliência entre data centers.
7. A. A opção A é a melhor resposta. Começar com um template existente, preencher parâmetros e gerar o comando gcloud é a maneira mais confiável. A opção D pode funcionar, mas vários parâmetros necessários para sua configuração podem não estar no script com o qual você começa. Pode haver alguma tentativa e erro com essa opção. As opções B e C podem levar a uma solução mas podem levar algum tempo para serem completadas.

8. A. O comando correto é a opção A. A opção B tem size em vez de num-nodes. A opção C tem region-nodes em vez de num-nodes. A opção D está faltando o nome do parâmetro --num-nodes.
9. C. Time to Live não é um atributo de deployments, então a opção C é a resposta correta. Nome da aplicação, imagem do contêiner e comando inicial podem todos ser especificados.
10. B. Arquivos de configuração de deployment criados no Cloud Console são salvos no formato YAML. CSV, TSV e JSON não são usados.
11. C. O comando kubectl é usado para controlar cargas de trabalho em um cluster Kubernetes uma vez que ele é criado, então a opção C está correta. As opções A e B estão incorretas porque gcloud não é usado para manipular processos do Kubernetes. A opção D está errada porque container não é necessário nos comandos kubectl.
12. C. A opção C é o comando correto. A opção A usa o termo upgrade em vez de scale. A opção B usa incorretamente gcloud. A opção D usa o parâmetro incorreto pods.
13. D. O Cloud Operations Suite é um serviço abrangente de monitoramento, log, alerta e notificação que pode ser usado para monitorar clusters Kubernetes.
14. D. O GKE envia métricas e logs para o Cloud Monitoring e o Cloud Logging por padrão, então você não precisa fazer nada além de aceitar a configuração padrão para monitoramento e log. Não existem parâmetros --monitoring=True e --logging=True. Pools de nós são usados para agrupar nós com configurações semelhantes e não são necessários para monitoramento e log. Namespaces são usados para separar logicamente cargas de trabalho em clusters e não precisam ser configurados individualmente para habilitar monitoramento e log.
15. A. Prometheus é uma ferramenta de monitoramento de código aberto popular disponível como um serviço gerenciado no Google Cloud. O Apache Flink é uma plataforma de processamento de stream e lote semelhante ao Cloud Dataflow. O MongoDB é um banco de dados NoSQL que usa um modelo de armazenamento de documentos. O Spark é uma ferramenta de análise de dados disponível como um serviço gerenciado no Google Cloud, mas não é uma ferramenta de monitoramento.
16. B. Clusters no modo Autopilot requerem a menor configuração e gestão de infraestrutura, então B é a resposta correta. Clusters no modo Padrão requerem que você especifique opções de infraestrutura e configuração. As opções C e D são modos fictícios de clusters.
17. A. Clusters no modo Padrão requerem que você faça escolhas de configuração e infraestrutura, então A é a resposta correta. Clusters no modo Autopilot usam infraestrutura otimizada pré-configurada e não dão tanto controle sobre configuração e infraestrutura quanto o modo Padrão. As opções C e D são modos fictícios de clusters.

18. B. B é a resposta correta porque com uma configuração de canal estático, o GKE não atualizará automaticamente o cluster. A opção A é a escolha correta se você deseja atualizações automáticas. Pools de nós e ReplicaSets não estão relacionados a configurações de atualização.
19. A. Todas as interações com o cluster são feitas através do mestre usando a API do Kubernetes. Se uma ação deve ser tomada em um nó, o comando é emitido pelo plano de controle, então a opção A é a resposta correta. As opções B e D estão incorretas porque são controladores dentro do cluster e não impactam como comandos são recebidos de dispositivos clientes. A opção C está incorreta porque kubectl, não gcloud, é usado para iniciar deployments.
20. A. Serviços fornecem um nível de indireção para acessar pods. Pods são efêmeros. Clientes se conectam a serviços, que podem descobrir pods. ReplicaSets e StatefulSets fornecem pods gerenciados. Alertas são para relatar o estado dos recursos.

Capítulo 8: Gerenciando Clusters Kubernetes no Modo Padrão

1. B. Quando nas páginas do Cloud Console, você pode clicar no nome do cluster para ver a página de Detalhes, então a opção B é a resposta correta. Digitar o nome de um cluster na barra de pesquisa nem sempre retorna os detalhes do cluster; pode retornar detalhes do grupo de instâncias. Não existe um comando como `gcloud cluster details`.
2. A. Você pode encontrar o número de vCPUs na listagem do cluster na seção Node Pools da página de Detalhes dos Nós. As outras seções não têm detalhes sobre vCPUs.
3. B. O comando correto inclui `gcloud container` para descrever o serviço, `clusters` para indicar o recurso ao qual você está se referindo, e `list` para indicar o comando, o que torna a opção B a resposta correta. As opções A e C não são comandos válidos.
4. B. É provável que você não tenha privilégios de acesso ao cluster. O comando `gdcloud container clusters get-credentials` é o comando correto para configurar o `kubectl` para usar as credenciais do Google Cloud para o cluster, então a opção B é a correta. As opções A, C e D são comandos inválidos.
5. C. Clicar no botão Editar permite mudar, adicionar ou remover rótulos, então a opção C é a resposta correta. O botão Conectar está na página de listagem do cluster, e o botão Implantar é para criar novos deployments. Não há como entrar rótulos na seção Rótulos ao exibir detalhes.
6. D. Ao redimensionar, o comando `gcloud container clusters resize` exige o nome do cluster e do node pool a modificar. O tamanho é necessário para especificar quantos nós devem estar em execução. Portanto, a opção D está correta.
7. B. Pods são usados para implementar réplicas de um deployment. É uma boa prática modificar os deployments, que são configurados com uma especificação do número de réplicas que devem sempre estar em execução, então a opção B é a resposta correta. A opção A está incorreta; você não deve modificar pods diretamente. As opções C e D estão incorretas porque não alteram o número de pods executando uma aplicação.
8. C. Deployments são listados em Workloads, tornando a opção C a resposta correta. A opção Cluster mostra detalhes sobre clusters mas não tem detalhes sobre deployments. Storage mostra informações sobre volumes persistentes e classes de armazenamento. Deployments não é uma opção.
9. B. Existem quatro ações disponíveis para deployments (Autoscale, Expose, Rolling Update e Scale), então a opção B está correta. Adicionar, Modificar e Excluir não são opções.
10. C. Como os deployments são gerenciados pelo Kubernetes e não pelo Google Cloud, precisamos usar um comando `kubectl` e não um comando `gcloud`, o que torna a opção C correta. A opção D está incorreta porque segue a estrutura de

comando do gcloud, não a estrutura de comando do kubectl. O comando kubectl tem o verbo, como get, antes do tipo de recurso, como deployments, por exemplo.

11. D. Você pode especificar a imagem do contêiner, o nome do cluster e o nome da aplicação junto com os rótulos, comando inicial e namespace; portanto, a opção D é a resposta correta.
12. A. A página de Detalhes do Deployment inclui aplicações, então a opção A está correta. Contêineres são usados para implementar serviços; detalhes do serviço não estão disponíveis lá. A página de Detalhes do Cluster não contém informações sobre serviços em execução no cluster.
13. A. kubectl run é o comando usado para iniciar um deployment. Ele recebe um nome para o deployment, uma imagem e uma especificação de porta. As outras opções não são comandos kubectl válidos.
14. A. A opção A mostra o comando correto, que é kubectl delete service ml-classifier-3. A opção B está faltando o termo service. As opções C e D não podem estar corretas porque os serviços são gerenciados pelo Kubernetes, não pelo Google Cloud.
15. C. O Container Registry é o serviço para gerenciar imagens que podem ser usadas em outros serviços, incluindo o Kubernetes Engine e o Compute Engine, tornando a opção C correta. Tanto o Compute Engine quanto o Kubernetes Engine usam imagens mas não as gerenciam. Não existe um serviço chamado Container Engine.
16. A. As imagens são gerenciadas pelo Google Cloud, então o comando correto será um comando gcloud, tornando a opção A a resposta correta. A opção B está incorreta porque o verbo é colocado antes do recurso. As opções C e D estão incorretas porque kubectl é para gerenciar recursos do Kubernetes, não recursos do Google Cloud como imagens de contêiner.
17. B. O comando correto é gcloud container images describe, o que torna a opção B a resposta correta. describe é o verbo ou operação do gcloud para mostrar os detalhes de um objeto. Todas as outras opções são comandos inválidos.
18. B. O comando kubectl expose deployment torna um serviço acessível, então a opção B está correta. Endereços IP são atribuídos a VMs, não a serviços. O comando gcloud não gerencia serviços do Kubernetes, então a opção C está incorreta. A opção D está incorreta porque tornar um serviço acessível não é uma tarefa em nível de cluster.
19. B. O autoscaling é a maneira mais econômica e menos onerosa de responder a mudanças na demanda por um serviço, então a opção B é a resposta correta. A opção A pode executar nós mesmos quando não são necessários. A opção C é manualmente intensiva e requer intervenção humana. A opção D reduz a intervenção humana, mas não leva em conta picos ou quedas inesperadas na demanda.

20. B. Engenheiros de nuvem que trabalham com Kubernetes precisarão estar familiarizados com o trabalho com clusters, nós, pods e imagens de contêiner. Eles também precisarão estar familiarizados com deployment. A opção B é a resposta correta porque as outras opções estão todas faltando um componente importante do Kubernetes que os engenheiros de nuvem terão que gerenciar.

Capítulo 9: Computação com Cloud Run e App Engine

1. C. Serviços do Cloud Run são serviços gerenciados e sem servidor para a execução de contêineres, projetados para suportar contêineres que rodam continuamente, o que é necessário para um serviço de API. Compute Engine exige que você gerencie servidores, então as opções A e B estão incorretas. A opção D está incorreta porque os trabalhos do Cloud Run são para contêineres que realizam uma tarefa e depois terminam.
2. D. Trabalhos do Cloud Run são serviços gerenciados e sem servidor para a execução de contêineres, projetados para suportar executáveis que rodam até que uma tarefa seja concluída. Kubernetes Engine é usado para executar contêineres, mas é mais adequado para executar um grande número de contêineres em ambientes complicados, como ambientes que precisam suportar múltiplos namespaces. Compute Engine poderia ser usado, mas requer mais administração do que usar o Cloud Run. App Engine Flexível poderia ser usado para executar um contêiner, mas o Cloud Run é preferido ao App Engine.
3. B. Trabalhos de array no Cloud Run permitem que múltiplos contêineres sejam executados e processem a carga de trabalho em paralelo. Como os dados em cada arquivo são independentes dos dados em outros arquivos, eles podem ser processados em qualquer ordem e em paralelo. A opção A está incorreta porque os dados são publicamente disponíveis e não há necessidade de chaves de criptografia gerenciadas pelo cliente. A opção C está incorreta porque não há menção de uma necessidade de escrever dados em um banco de dados Cloud SQL. A opção D está incorreta; um endereço IP privado vs. público não é relevante para a questão.
4. C. Ao suportar a afinidade de sessão na configuração de Conexão, o Cloud Run roteará todas as solicitações de um cliente para o mesmo contêiner, se possível. A opção A está incorreta; Conexão Cloud SQL é usada para conectar um serviço do Cloud Run ao Cloud SQL. A opção B está incorreta; trabalhos de array no Cloud Run permitem que múltiplos contêineres sejam executados e processem a carga de trabalho em paralelo. A opção D está incorreta; um endereço IP privado vs. público não é relevante para a questão.
5. A. A configuração de ingresso interno restringirá o tráfego de rede ao tráfego interno do Google Cloud. A opção B está incorreta; isso permitiria tráfego que entra através de平衡amento de carga externo. A opção C está incorreta; isso permitiria todo o tráfego. A opção D está incorreta; não existe tal configuração para Tráfego Proxy PII no Cloud Run.
6. B. Você especifica uma conta de serviço para um serviço do Cloud Run na aba Segurança da página Criar Serviço no console do Cloud Run. A opção A está incorreta; isso é usado para configurar conexão de rede. A opção C está incorreta; isso é usado para configurar o contêiner em execução. A opção D está incorreta; isso é para definir variáveis de ambiente e referenciar segredos.
7. A. A resposta correta é A; você concederia uma função com permissões apropriadas a um grupo que inclui os desenvolvedores que precisam de acesso. A

opção B está incorreta; o Cloud Identity Aware Proxy garante que os usuários sejam autenticados e autorizados, mas não concede permissões. A opção C está incorreta; a política de ingresso controla o tráfego de rede, não usuários. A opção D está incorreta; a aba Segurança não concede acesso a usuários.

8. C. A resposta correta é C; uma Conexão VPC permite o uso do Acesso VPC Sem Servidor para conectar seu serviço do Cloud Run a outros recursos na sua VPC. A opção A está incorreta; isso é usado para conectar a um banco de dados Cloud SQL. A opção B está incorreta; o Proxy IAP é usado para autenticar e autorizar usando controles de acesso granulares. A opção D está incorreta; a afinidade de sessão é usada para enviar todas as solicitações de um cliente para o mesmo contêiner.
9. D. A resposta correta é D; configurar o serviço para usar HTTP/2 de ponta a ponta permitirá o uso do protocolo gRPC. A opção A está incorreta; suporte para tráfego de平衡amento de carga externo é configurado usando a configuração de ingresso e não é necessariamente requerido para usar gRPC. A opção B está incorreta; o Proxy IAP é usado para autenticar e autorizar usando controles de acesso granulares. A opção C está incorreta; a afinidade de sessão é usada para enviar todas as solicitações de um cliente para o mesmo contêiner.
10. B. A resposta correta é B; tanto o Container Registry quanto o Artifact Registry podem ser usados para armazenar e tornar imagens de contêiner acessíveis ao Cloud Run. As opções A e C estão ambas faltando uma opção válida para servir imagens de contêiner. A opção D está incorreta; Kubernetes é usado para orquestração de contêineres, mas não fornece serviços de registro de imagens de contêiner para o Cloud Run.
11. B. Versões suportam migração. Um aplicativo pode ter múltiplas versões, e ao implantar com o parâmetro --migrate, você pode migrar o tráfego para a nova versão, então a opção B é a resposta correta. Serviços são uma abstração de nível mais alto e representam a funcionalidade de um microsserviço. Um aplicativo pode ter vários serviços, mas eles servem a propósitos diferentes. Instâncias executam código em uma versão. Instâncias podem ser adicionadas e removidas conforme necessário, mas elas executarão apenas uma versão de um serviço. Grupos de instâncias fazem parte do Compute Engine e não são um componente do App Engine.
12. A. O autoscaling permite definir um número máximo e mínimo de instâncias, o que torna a opção A correta. O escalonamento básico não suporta instâncias máximas e mínimas. A opção C não é recomendada porque é difícil prever quando a carga atingirá o pico, e mesmo que o horário seja previsível hoje, pode mudar ao longo do tempo. A opção D está errada; não há opção de detecção de instância.
13. B. O comando correto é gcloud app deploy, o que faz da opção B a correta. As opções A e C estão incorretas porque os comandos gcloud components são usados para instalar comandos gcloud para trabalhar com partes do App Engine, como o ambiente de tempo de execução Python. A opção D está incorreta; você não precisa especificar uma instância no comando.

14. B. O arquivo app.yaml é usado para configurar um aplicativo App Engine, o que torna a opção B correta. As outras opções não são arquivos usados para configurar o App Engine.
15. A. max_concurrent_requests permite especificar o número máximo de solicitações concorrentes antes que outra instância seja iniciada, o que torna a opção A correta. target_throughput_utilization funciona de forma semelhante, mas usa uma escala de 0,05 a 0,95 para especificar a utilização máxima de throughput. max_instances especifica o número máximo de instâncias, mas não os critérios para adicionar instâncias. max_pending_latency é baseado no tempo que uma solicitação espera, não no número de solicitações.
16. C. O escalonamento básico permite apenas tempo ocioso e instâncias máximas, então a opção C é a resposta correta. min_instances não é suportado. target_throughput_utilization é um parâmetro de autoscaling, não um parâmetro de escalonamento básico.
17. C. O parâmetro runtime especifica o ambiente de linguagem para execução, o que torna a opção C correta. O script a ser executado é especificado pelo parâmetro script. A URL para acessar o aplicativo é baseada no nome do projeto e no domínio appspot.com. Não há parâmetro para especificar o tempo máximo que um aplicativo pode rodar.
18. A. Usar instâncias dinâmicas especificando autoscaling ou escalonamento básico ajustará automaticamente o número de instâncias em uso com base na carga, então a opção A está correta. A opção B está incorreta porque autoscaling e escalonamento básico criam apenas instâncias dinâmicas. As opções C e D estão incorretas porque o escalonamento manual não ajustará instâncias automaticamente, então você pode continuar executando mais instâncias do que necessário em alguns momentos.
19. B. --split-by é o parâmetro usado para especificar o método de divisão de tráfego. Opções válidas são cookie, ip e random. Todas as outras opções não são parâmetros válidos para o comando gcloud app services set-traffic.
20. D. Todos os três métodos listados, endereço IP, cookie HTTP e divisão aleatória, são métodos permitidos para divisão de tráfego.

Capítulo 10: Computing with Cloud

1. C. Cloud Run é um serviço sem servidor para executar aplicações containerizadas que funcionam continuamente e fornecem um endpoint, tornando a opção C a resposta correta. Isso é diferente de Cloud Functions, que é projetado para suportar funções de propósito único que operam de forma independente e em resposta a eventos isolados no Google Cloud e completam dentro de um período de tempo especificado. Compute Engine não é uma opção sem servidor. Cloud Storage não é um produto de computação.
2. C. Um período de timeout muito baixo explicaria por que os arquivos menores são processados a tempo, mas os maiores não, o que torna a opção C a resposta certa. Se apenas 10% dos arquivos estão falhando, então não é um erro de sintaxe ou a escolha errada do runtime, como nas opções A e B. Esses erros afetariam todos os arquivos, não apenas os maiores. Da mesma forma, se houvesse um problema de permissão com o bucket do Cloud Storage, afetaria todos os arquivos.
3. B. Essas ações são conhecidas como eventos na terminologia do Google Cloud; portanto, a opção B é a resposta correta. Um incidente pode ser um ocorrido relacionado à segurança ou ao desempenho, mas esses são não relacionados às ações esperadas e padronizadas que constituem eventos. Um gatilho é uma declaração de que uma certa função deve executar quando um evento ocorre. Uma entrada de log está relacionada às aplicações registrando dados sobre eventos significativos. As entradas de log são úteis para monitoramento e conformidade, mas em si não são ações relacionadas a eventos.
4. C. A resposta correta é a opção C porque SSL é um protocolo seguro para acessar servidores remotamente. É usado, por exemplo, para acessar instâncias no Compute Engine. Ele não tem eventos que podem ser acionados usando Cloud Functions. Os três produtos GCP listados geram eventos que podem ter gatilhos associados a eles.
5. D. A resposta correta é D; todas as outras opções estão faltando dois ou mais ambientes suportados. Os ambientes de runtime suportados incluem Node.js, Python, Go, Java, .NET, Ruby e PHP.
6. C. Solicitações HTTP usando GET, POST, DELETE, PUT e OPTIONS podem invocar um gatilho HTTP em Cloud Functions, então a opção C é a resposta certa.
7. D. A resposta correta, opção D, mostra os quatro eventos suportados no Cloud Storage:
 - google.storage.object.finalize
 - google.storage.object.delete
 - google.storage.object.archive
 - google.storage.object.metadataUpdate
8. C. Não há opção para especificar o tipo de arquivo ao qual a função se aplica, então a opção C está correta. No entanto, você pode especificar o bucket ao qual

a função é aplicada. Você poderia apenas salvar arquivos ou os tipos que deseja processar naquele bucket, ou você poderia fazer com que sua função verifique o tipo de arquivo e então execute o resto da função ou não, baseado no tipo. Todas as outras opções listadas são parâmetros para uma função do Cloud Storage.

9. D. Cloud Functions de Segunda Geração podem ter entre 128 MB e 16 GB de memória alocados, o que torna a opção D a resposta correta.
10. B. Por padrão, Cloud Functions pode executar por até 1 minuto antes de esgotar o tempo, então a opção B está correta. No entanto, você pode definir o parâmetro de timeout para uma Cloud Function por períodos de até 9 minutos antes de esgotar o tempo.
11. A. A opção A instalará comandos padrão do gcloud. As opções B, C e D não são comandos válidos do gcloud.
12. A. O gatilho correto na opção A é google.storage.object.finalize, que ocorre após um arquivo ser carregado. A opção B não é um nome de gatilho válido. A opção C é acionada quando um arquivo é arquivado, não carregado. A opção D é acionada quando algum atributo de metadado muda, mas não necessariamente apenas após um arquivo ser carregado.
13. C. Os três parâmetros são runtime, trigger-resource e trigger-event, como listado na opção C. Todos devem ser definidos, então as opções A e B estão incorretas. file-type não é um parâmetro para criar uma Cloud Function no Cloud Storage, então a opção D está incorreta.
14. A. A resposta correta é a opção A, gcloud functions delete. A opção B faz referência a componentes, o que está incorreto. Você precisa fazer referência a componentes ao instalar ou atualizar comandos do gcloud, mas não ao excluir uma função da nuvem, então as opções B e C estão incorretas. A opção D está incorreta porque o tipo de entidade do Google Cloud, neste caso funções, vem antes do nome da operação, neste caso delete, em um comando gcloud.
15. B. As mensagens são armazenadas em um formato de texto, base64, para que dados binários possam ser armazenados na mensagem em um formato de texto, então a opção B está correta. A opção A está incorreta; é necessário para mapear de uma codificação binária para uma codificação de texto padrão. A opção C está incorreta porque a função não preenche com caracteres extras para torná-los do mesmo comprimento. A opção D está incorreta; ela não muda tipos de dados de dicionário para tipos de dados de lista.
16. C. A opção C está correta porque inclui o nome da função, o ambiente de runtime e o nome do tópico Pub/Sub. A opção A está incorreta porque falta tanto o runtime quanto o tópico. A opção B está incorreta porque falta o tópico. A opção D está incorreta porque a especificação do runtime está incorreta; você tem que especificar python37 e não python como o runtime.
17. B. Há apenas um tipo de evento que é acionado no Cloud Pub/Sub, e isso é quando uma mensagem é publicada, o que torna a opção B a resposta correta. A opção A está incorreta; o Cloud Pub/Sub tem um tipo de evento que pode ter um gatilho.

A opção C está incorreta; o Cloud Pub/Sub não analisa o código para determinar quando ele deve ser executado. A opção D está incorreta; você não precisa especificar um tipo de evento com funções do Cloud Pub/Sub.

18. B. A resposta correta é a opção B porque utiliza um evento de finalização do Cloud Storage para acionar a conversão, se necessário. Há um atraso mínimo entre o momento em que o arquivo é carregado e quando é convertido. A opção A é uma possibilidade, mas exigiria mais codificação do que a opção B. A opção C não é uma boa opção porque os arquivos não são convertidos até que o trabalho em lote seja executado. A opção D está incorreta porque você não pode criar uma Cloud Function para o Cloud Pub/Sub usando um evento de finalização. Esse evento é para o Cloud Storage, não para o Cloud Pub/Sub.
19. D. Todas as opções estão disponíveis junto com zip do Cloud Storage.
20. A. O gatilho HTTP permite o uso de chamadas POST, GET e PUT, então a opção A é a resposta correta. Webhook e Cloud HTTP não são tipos de gatilho válidos.

Capítulo 11: Planejamento de Armazenamento na Nuvem

1. D. Uma classe de armazenamento Arquivo não pode ser alterada para padrão ou para qualquer outra classe de armazenamento. Todas as outras opções são permitidas.
2. C. O objetivo é reduzir custos, então você gostaria de usar a opção de armazenamento menos custosa. Coldline é projetado para objetos que são acessados no máximo uma vez a cada 90 dias, então a opção C está correta. Nearline e padrão custam mais do que coldline, então essas não são boas opções. Arquivo deve ser usado apenas para objetos acessados no máximo uma vez por ano.
3. B. Bigtable é um banco de dados de colunas largas que pode ingerir grandes volumes de dados consistentemente, então a opção B está correta. Ele também suporta latência de milissegundos baixa, tornando-o uma boa escolha para suportar consultas. Cloud Spanner é um banco de dados relacional global que não é adequado para a ingestão rápida de grandes volumes de dados. Firestore é um modelo de dados de objeto e não é uma boa opção para dados de IoT ou outros dados de séries temporais. BigQuery é um banco de dados analítico e não é projetado para a ingestão de grandes volumes de dados com latências de escrita curtas.
4. A. A opção A está correta porque o Memorystore é um cache gerenciado. O cache pode ser usado para armazenar os resultados de consultas. Consultas subsequentes que referenciam os dados armazenados no cache podem lê-lo do cache, o que é muito mais rápido do que ler de discos persistentes. SSDs têm latência significativamente menor do que discos rígidos e devem ser usados para aplicações sensíveis ao desempenho, como bancos de dados. As opções B e D estão incorretas porque discos persistentes HDD não oferecem o melhor desempenho em termos de IOPS. As opções C e D estão incorretas porque o Firestore é um banco de dados NoSQL gerenciado e não atenderia à exigência de continuar usando um banco de dados relacional.
5. B. HDDs são a melhor escolha para discos persistentes para um banco de dados local quando o desempenho não é a principal preocupação e você está tentando manter os custos baixos, então a opção B está correta. A opção A está errada porque SSDs são mais caros e os usuários não precisam da menor latência disponível. As opções C e D estão erradas; ambos são outros bancos de dados que não seriam usados para armazenar dados em um banco de dados relacional local.
6. B. Configurações de ciclo de vida não podem mudar a classe de armazenamento de arquivo para padrão, então a opção B é a resposta certa. A opção A é verdadeira; você pode definir períodos de retenção ao criar um bucket. A opção C é verdadeira; o Cloud Storage não fornece acesso semelhante a um sistema de arquivos aos blocos de dados internos. A opção D é verdadeira porque o Cloud Storage é altamente durável.

7. A. A versão mais recente de um objeto é chamada de versão ativa, então a opção A está correta. As opções B, C e D estão incorretas; top e active não são termos usados para se referir a versões.
8. B. Tanto o Cloud SQL quanto o Spanner são bancos de dados relacionais e são bem adequados para aplicações de processamento de transações, então a opção B está correta. A opção A está incorreta porque o BigQuery é um banco de dados analítico projetado para data warehousing e análises, não para processamento de transações. As opções C e D estão incorretas porque o Bigtable é um banco de dados NoSQL de colunas largas, não um banco de dados relacional.
9. C. Tanto o MySQL quanto o PostgreSQL são opções do Cloud SQL, então a opção C está correta. As opções A e B estão incorretas; Oracle não é uma opção do Cloud SQL. A opção D está incorreta porque DB2 não é uma opção do Cloud SQL. Você poderia escolher executar DB2 ou Oracle em suas instâncias, mas teria que gerenciá-los, ao contrário dos bancos de dados gerenciados do Cloud SQL.
10. D. A localização multirregional e multi-super-regional de nam-eur-asia1 é a mais cara, o que torna a opção D a resposta certa. A opção A é uma região que custa menos do que a multi-super-regional nam-eur-asia1. A opção C está incorreta; essa é uma zona, e o Spanner é configurado para regiões ou super regiões. A opção B está incorreta; é apenas uma única super região, o que custa menos do que a implantação em múltiplas super regiões.
11. D. O BigQuery e o Firestore são todos serviços totalmente gerenciados que não requerem que você especifique informações de configuração para VMs, o que torna a opção D correta. O Cloud SQL e o Bigtable exigem que você especifique algumas informações de configuração para VMs.
12. B. O Firestore é um banco de dados de documentos, o que torna a opção B correta. O Cloud SQL e o Spanner são bancos de dados relacionais. O Bigtable é um banco de dados de colunas largas. O Google não oferece um banco de dados de grafos gerenciado.
13. A. O BigQuery é um serviço gerenciado projetado para data warehouses e análises. Ele usa SQL padrão para consultas, o que torna a opção A a resposta correta. O Bigtable pode suportar o volume de dados descrito, mas não usa SQL como linguagem de consulta. O Cloud SQL não é a melhor opção para escalar para dezenas de petabytes. O IBM DB2 é um banco de dados relacional, mas não é um serviço de banco de dados gerenciado pelo Google Cloud.
14. B. O Firestore é um banco de dados de documentos que possui recursos de suporte móvel, como sincronização de dados, então a opção B está correta. O BigQuery é para análises, não para aplicações móveis ou transacionais. O Spanner é um banco de dados relacional global, mas não possui recursos específicos para móveis. O Bigtable poderia ser usado com dispositivos móveis, mas não possui recursos específicos para móveis como sincronização.
15. D. Além dos padrões de leitura e escrita, custo e consistência, você deve considerar suporte a transações e latência, o que torna a opção D correta.

16. B. A opção B está correta porque o Memorystore pode ser configurado para usar entre 1 GB e 300 GB de memória.
17. D. Uma vez que um bucket é definido para arquivo, ele não pode ser alterado para outra classe de armazenamento; assim, a opção D está correta. Padrão pode mudar para Nearline, Coldline ou Arquivo. Nearline pode mudar para Coldline e Arquivo. Coldline pode mudar para Arquivo.
18. A. Para usar o BigQuery para armazenar dados, você deve ter um conjunto de dados para armazená-lo, o que torna a opção A a resposta correta. Buckets são usados pelo Cloud Storage, não pelo BigQuery. Você não gerencia discos persistentes ao usar o BigQuery. Uma entidade é uma estrutura de dados no Firestore, não no BigQuery.
19. D. Com um banco de dados MySQL, você pode configurar a versão do MySQL, conectividade, tipo de máquina, backups automáticos, réplicas de failover, flags de banco de dados, janelas de manutenção e rótulos, então a opção D está correta.
20. A. Taxas de acesso são usadas com armazenamento Nearline e Coldline, o que torna a opção A correta. Não há cobrança de transferência envolvida. A opção C está incorreta porque as taxas de egresso seriam aplicadas antes da mudança para Nearline e Coldline. A opção D está incorreta porque nearline e coldline incorrem em taxas de acesso.

Capítulo 12: Implementando Armazenamento no Google Cloud

1. C. Criar bancos de dados é responsabilidade dos administradores de banco de dados ou outros usuários do Cloud SQL, então a opção C está correta. O Google aplica patches de segurança e realiza outras manutenções, então a opção A está incorreta. O Google Cloud realiza backups programados regularmente, então a opção B está incorreta. Administradores de banco de dados precisam agendar backups, mas o Google Cloud garante que eles sejam realizados conforme o programado. Usuários do Cloud SQL não podem usar SSH para conectar-se a um servidor Cloud SQL, então eles não podem ajustar o sistema operacional. Isso não é um problema; o Google cuida disso.
2. A. Cloud SQL é controlado usando o comando gcloud; a sequência de termos nos comandos gcloud é gcloud seguido pelo serviço, neste caso SQL; seguido por um recurso, neste caso backups; e um comando ou verbo, neste caso create. A opção A é a resposta correta. A opção B está incorreta porque gsutil é usado para trabalhar com Cloud Storage, não Cloud SQL. A opção C está errada porque a ordem dos termos está incorreta; backups vem antes de create. A opção D está incorreta porque o comando ou verbo deveria ser create.
3. A. A opção A é a resposta correta. O comando base é gcloud sql instances patch, que é seguido pelo nome da instância e um horário de início passado para o parâmetro --backup-start-time. A opção B está incorreta porque databases não é o recurso correto para referenciar; instances é. A opção C usa o comando cbt, que é para uso com Bigtable, então está incorreta. Da mesma forma, a opção D está incorreta porque usa o comando bq, que é usado para gerenciar recursos do BigQuery.
4. C. O modo Datastore usa uma linguagem de consulta semelhante ao SQL chamada GQL, então a opção C está correta. A opção A está incorreta; SQL não é usado com esse banco de dados. A opção B está incorreta; MDX é uma linguagem de consulta para sistemas de processamento analítico online (OLAP). A opção D está incorreta porque DataFrames é uma estrutura de dados usada no Spark.
5. C. A opção C é o comando correto. Ele tem o comando base correto, gcloud firestore export, seguido pelo nome de um bucket do Cloud Storage para conter o arquivo de exportação. A opção A está incorreta porque o nome do parâmetro collection não é necessário. A opção B está incorreta porque dump não é uma operação válida e o termo collection não é necessário. A opção D está incorreta porque usa o comando ou verbo dump em vez de export.
6. C. A opção C está correta; o BigQuery exibe uma estimativa da quantidade de dados escaneados. Isso é importante porque o BigQuery cobra pelos dados escaneados nas consultas. A opção A está incorreta; saber quanto tempo você levou para inserir uma consulta não é útil. A opção B está incorreta; você precisa usar a estimativa de dados escaneados com a Calculadora de Preços para obter uma estimativa de custo. A opção D está incorreta; você não cria clusters no BigQuery como faz com o Bigtable e o Dataproc. Dados de E/S de rede não são exibidos.

7. B. A opção B mostra a estrutura de comando bq correta, que inclui localização e a opção --dry_run. Esta opção calcula uma estimativa sem realmente executar a consulta. As opções A e C estão incorretas porque usam o comando errado; gcloud e gsutil não são usados com o BigQuery. A opção D também está errada. cbt é uma ferramenta para trabalhar com Bigtable, não BigQuery. Tenha cuidado para não confundir os dois, pois seus nomes são semelhantes.
8. A. A opção A está correta; a opção de menu é Histórico Pessoal ou Histórico do Projeto. As opções B e C estão incorretas; não existe a opção Trabalhos Ativos ou Meus Trabalhos. Histórico de Trabalhos mostra trabalhos ativos, trabalhos concluídos e trabalhos que geraram erros. A opção D está incorreta; você pode obter o status do trabalho no console.
9. C. O BigQuery fornece uma estimativa da quantidade de dados escaneados, e a Calculadora de Preços dá uma estimativa de custo para escanear esse volume de dados. As opções A, B e D estão incorretas; o serviço de Faturamento rastreia as cobranças incorridas - não é usado para estimar cobranças futuras ou potenciais.
10. B. A opção B está correta; o próximo passo é criar um banco de dados dentro da instância. Uma vez que um banco de dados é criado, tabelas podem ser criadas, e dados podem ser carregados nas tabelas. A opção A está incorreta; o Cloud Spanner é um banco de dados gerenciado, então você não precisa aplicar patches de segurança. A opção C está incorreta porque você não pode criar tabelas sem ter criado um banco de dados primeiro. A opção D está incorreta; nenhuma tabela é criada na qual você poderia importar dados quando uma instância é criada.
11. D. A opção D está correta porque não é necessário aplicar patches nos recursos de computação subjacentes ao usar o Cloud Spanner, pois o Google gerencia os recursos usados pelo Cloud Spanner. Atualizar pacotes é uma boa prática ao usar VMs, por exemplo, com o Compute Engine, mas não é necessário com um serviço gerenciado.
12. C. Este caso de uso é bem adequado para o Pub/Sub, então a opção C está correta. Envolve enviar mensagens para o tópico, e o modelo de assinatura é um bom ajuste. O Pub/Sub tem um período de retenção para suportar o período de retenção de três dias. A opção A está incorreta; o Bigtable é projetado para armazenar grandes volumes de dados. O Dataproc é para processar e analisar dados, não para passá-los entre sistemas. O Cloud Spanner é um banco de dados relacional global. Você poderia projetar um aplicativo para atender a este caso de uso, mas exigiria um desenvolvimento substancial e seria caro para executar.
13. C. O Pub/Sub trabalha com tópicos, que recebem e mantêm mensagens, e assinaturas, que disponibilizam mensagens para aplicativos consumidores; portanto, a opção C está correta. A opção A está incorreta; tabelas são estruturas de dados em bancos de dados relacionais, não filas de mensagens. Da mesma forma, a opção B está errada porque bancos de dados existem em instâncias de sistemas de gerenciamento de bancos de dados, não sistemas de mensagens. A opção D está errada porque tabelas não são um recurso em sistemas de mensagens.

14. C. O comando correto é gcloud components install cbt para instalar a ferramenta de linha de comando do Bigtable, então a opção C está correta. As opções A e B estão incorretas; apt-get é usado para instalar pacotes em alguns sistemas Linux, mas não é específico para o Google Cloud. A opção D está incorreta; não existe tal comando como bigtable-tools.
15. A. Você precisaria usar um comando cbt, que é a ferramenta de linha de comando para trabalhar com o Bigtable, então a opção A está correta. Todas as outras opções referenciam gcloud e, portanto, estão incorretas.
16. B. O Cloud Dataproc é um serviço gerenciado para Spark e Hadoop, então a opção B está correta. Cassandra é um banco de dados distribuído de big data, mas não é oferecido como um serviço gerenciado pelo Google, então as opções A e C estão incorretas. A opção D está incorreta porque o TensorFlow é uma plataforma de aprendizado profundo não incluída no Dataproc.
17. B. O comando correto é gcloud dataproc clusters create seguido pelo nome do cluster e o parâmetro --zone, então a opção B está correta. A opção A está incorreta porque bq é a ferramenta de linha de comando para o BigQuery, não para o Dataproc. A opção C é um comando gcloud faltando um verbo ou comando, então não cria um cluster.
18. B. gsutil é o comando correto, então a opção B está correta. A opção A está incorreta porque comandos gcloud não são usados para gerenciar o Cloud Storage. Da mesma forma, as opções C e D estão incorretas porque cbt é usado para trabalhar com o Bigtable e bq é usado para trabalhar com o BigQuery.
19. B. O comando na opção B renomeia corretamente um objeto de um nome antigo para um novo. A opção A está incorreta porque usa um comando cp em vez de mv. A opção C não inclui nomes de buckets, então está incorreta. A opção D usa gcloud, mas gsutil é a ferramenta de linha de comando para trabalhar com o Cloud Storage.
20. A. O Dataproc com Spark e sua biblioteca de aprendizado de máquina são ideais para este caso de uso, então a opção A está correta. A opção B sugere Hadoop, mas não é uma boa escolha para aplicações de aprendizado de máquina. A opção C está incorreta porque o Spanner é projetado como um banco de dados relacional global com suporte para sistemas de processamento de transações, não sistemas analíticos e de aprendizado de máquina. A opção D está incorreta; SQL é uma linguagem de consulta poderosa, mas não suporta os tipos de algoritmos de aprendizado de máquina necessários para resolver o problema proposto.

Capítulo 13: Carregando Dados para Armazenamento

1. C. O gsutil é a ferramenta de linha de comando para trabalhar com o Cloud Storage. A opção C é a resposta correta porque mb, abreviação de "make bucket" (criar bucket), é o verbo que segue o gsutil para criar um bucket. A opção D está errada porque não é um comando completo. A opção A está incorreta porque create e buckets estão na ordem errada. O comando gcloud storage buckets create também poderia ser usado para criar um bucket. A opção B está errada porque usa o gsutil com uma sintaxe de comando usada pelo gcloud.
2. B. A resposta correta é a opção B; o gsutil é o comando para copiar arquivos para o Cloud Storage. A opção A está incorreta; o verbo é cp, não copy. A opção C está incorreta porque gcloud cp não é um comando completo. A opção D não é um comando válido do gcloud storage.
3. C. A partir do console, você pode fazer upload de arquivos e pastas. As opções A e B estão incorretas porque estão faltando uma operação que pode ser realizada no console. A opção D está incorreta porque não há operação diff no Cloud Console.
4. D. Ao exportar um banco de dados do Cloud SQL, as opções de formato de arquivo de exportação são CSV e SQL, o que torna a opção D correta. A opção A está incorreta porque XML não é uma opção. As opções B e C estão incorretas porque JSON não é uma opção.
5. A. A opção A, formato SQL, exporta um banco de dados como uma série de comandos de definição de dados SQL. Esses comandos podem ser executados em outro banco de dados relacional sem ter que criar um esquema primeiro. A opção B poderia ser usada, mas isso exigiria mapear colunas para colunas em um esquema que foi criado antes de carregar o CSV, e o administrador do banco de dados gostaria de evitar isso. As opções C e D estão incorretas porque não são opções de formato de arquivo de exportação.
6. C. A opção C é o comando correto, gcloud sql export sql, indicando que o serviço é o Cloud SQL, a operação é exportar, e o formato do arquivo de exportação é SQL. O nome do arquivo e o bucket de destino estão corretamente formados. A opção A está incorreta porque refere-se ao gcloud storage, não ao gcloud sql. A opção B está incorreta porque está faltando um parâmetro de formato de arquivo de exportação. A opção D está incorreta porque o nome do bucket e o nome do arquivo estão na ordem errada.
7. B. A opção B está correta porque XML não é uma opção no processo de exportação do BigQuery. Todas as outras opções estão disponíveis.
8. D. A opção D está correta porque YAML não é um formato de armazenamento de arquivo; é usado para especificar dados de configuração. As opções A, B e C são todos tipos de arquivo de importação suportados.
9. A. O comando correto é bq load na opção A. Os parâmetros autodetect e source_format e o caminho para a fonte estão corretamente especificados em

todas as opções. A opção B está incorreta porque usa o termo import em vez de load. As opções C e D estão incorretas porque usam gcloud em vez de bq.

10. B. A resposta correta é B porque o Dataflow é um serviço de pipeline para processar dados de streaming e em lote que implementa workflows usados pelo Cloud Spanner. A opção A está incorreta; o Dataproc é um serviço gerenciado de Hadoop e Spark, que é usado para análise de dados. A opção C está incorreta; o Firestore é um banco de dados NoSQL. A opção D está incorreta porque o bq é usado com o BigQuery apenas.
11. A. Dados do Bigtable são exportados usando um programa Java compilado, então a opção A está correta. A opção B está incorreta; não existe um comando gcloud bigtable. A opção C está incorreta; o bq não é usado com o Bigtable. A opção D está incorreta porque não exporta dados do Bigtable.
12. C. Exportar do Dataproc exporta dados sobre a configuração do cluster, o que torna a opção C correta. A opção A está incorreta; dados em DataFrames não são exportados. A opção B está incorreta; o Spark não possui tabelas para armazenar dados persistentemente como bancos de dados relacionais. A opção D está incorreta; nenhum dado do Hadoop é exportado.
13. C. A resposta correta é a opção C; o serviço Dataproc suporta o Apache Spark, que possui bibliotecas para aprendizado de máquina. As opções A e B estão incorretas; nenhuma é um serviço de análise ou aprendizado de máquina. A opção D, DataAnalyze, não é um serviço real.
14. A. A opção A mostra o comando correto, que usa o gcloud seguido pelo serviço, neste caso pubsub, seguido pelo recurso, neste caso topics; e finalmente o verbo, neste caso create. A opção B está incorreta porque os dois últimos termos estão fora de ordem. As opções C e D estão incorretas porque não usam o gcloud. O bq é a ferramenta de linha de comando para o BigQuery. O cbt é a ferramenta de linha de comando para o Bigtable.
15. C. A resposta correta, opção C, usa gcloud pubsub subscriptions create seguido pelo tópico e o nome da assinatura. A opção A está incorreta porque está faltando o termo subscriptions. A opção B está incorreta porque está faltando o nome da assinatura. A opção D está incorreta porque usa o gsutil em vez do gcloud.
16. B. Usar uma fila de mensagens entre serviços desacopla os serviços, então se um atrasa não faz com que outros serviços atrasem, o que torna a opção B correta. A opção A está incorreta porque adicionar uma fila de mensagens não mitiga diretamente quaisquer riscos de segurança que possam existir no sistema distribuído, como permissões excessivamente permissivas. A opção C está incorreta; adicionar uma fila não está diretamente relacionado a linguagens de programação. A opção D está incorreta; por padrão, filas de mensagens têm um período de retenção.
17. B. A resposta correta é B; gcloud components seguido por install e então beta. A opção A está incorreta porque beta e install estão na ordem errada. As opções C e D estão erradas porque usam commands em vez de components.

18. A. O nome correto do parâmetro é autodetect, o que torna a opção A correta. As opções B e C não são realmente parâmetros válidos do bq. A opção D é um parâmetro válido, mas retorna o tamanho estimado dos dados escaneados ao executar uma consulta.
19. A. Avro suporta compressão Deflate e Snappy. CSV suporta Gzip e sem compressão. XML e Thrift não são opções de formato de arquivo de exportação.
20. A resposta correta é A. Você incluiria o parâmetro auto-ack no comando gcloud pubsub subscriptions pull. A opção B está incorreta, você puxa de uma assinatura, não de um tópico. A opção C está incorreta, você não usa o gsutil para trabalhar com o Pub/Sub e with-acknowledgement não é um parâmetro válido. A opção D está incorreta porque with-acknowledgement não é um parâmetro válido.

Capítulo 14: Redes na Nuvem: Nuvens Privadas Virtuais e Redes Privadas Virtuais

1. D. Nuvens privadas virtuais são globais, então a opção D está correta. Por padrão, elas têm sub-redes em todas as regiões. Recursos em qualquer região podem ser acessados através da VPC. As opções A, B e C estão todas incorretas.
2. B. Faixas de IP são atribuídas a sub-redes, então a opção B está correta. Cada sub-rede é atribuída uma faixa de IP para seu uso exclusivo. Faixas de IP são atribuídas a estruturas de rede, não a zonas e regiões. VPCs podem ter múltiplas sub-redes, mas cada sub-rede tem sua própria faixa de endereço.
3. B. A opção B está correta; roteamento dinâmico é o parâmetro que especifica se as rotas são aprendidas regionalmente ou globalmente. A opção A está incorreta; DNS é um serviço de resolução de nomes e não está envolvido com roteamento. A opção C está incorreta; não existe um parâmetro de política de roteamento estático. A opção D está incorreta porque roteamento sistemático não é uma opção real.
4. A. A resposta correta é `gcloud compute networks create`, que é a opção A. A opção B está incorreta; `networks vpc` não é uma parte correta do comando. A opção C está incorreta porque `gsutil` é o comando usado para trabalhar com o Cloud Storage. A opção D está incorreta porque isso não existe.
5. A. A opção de Log de Fluxo do comando `create vpc` determina se os logs são enviados para o Cloud Logging, então a opção A está correta. A opção B, Acesso a IP Privado, determina se uma VM precisa de um endereço IP externo para usar serviços do Google. A opção C está incorreta porque Cloud Logging é o serviço, não um parâmetro usado ao criar uma sub-rede. A opção D está incorreta porque máscara de sub-rede de comprimento variável tem a ver com endereços CIDR, não com log.
6. C. VPCs compartilhadas podem ser criadas no nível da organização ou da pasta na hierarquia de recursos, então a opção C está correta. As opções A e B estão incorretas; VPCs compartilhadas não são criadas nos níveis de recurso ou projeto. A opção D está incorreta; VPCs compartilhadas não são aplicadas em sub-redes, que são recursos na hierarquia de recursos.
7. A. A resposta correta é a aba de Redes da seção Gerenciamento, Segurança, Discos, Redes, Sole Tenancy da página, o que torna a opção A correta. A aba Gerenciamento não é sobre configurações de sub-rede. A opção D está incorreta porque não leva às opções de Sole Tenancy.
8. A. O emparelhamento de redes VPC é usado para comunicações interprojetos. A opção B está incorreta; não existe emparelhamento interprojetos. As opções C e D estão incorretas; elas têm a ver com a ligação de redes locais com redes no Google Cloud.
9. B. O alvo pode ser todas as instâncias em uma rede, instâncias com tags de rede ou instâncias usando uma conta de serviço específica, então a opção B está correta.

A opção A está incorreta; Ação é ou Permitir ou Negar. A opção C está incorreta; Prioridade determina qual de todas as regras correspondentes é aplicada. A opção D está incorreta; ela especifica se a regra é aplicada ao tráfego de entrada ou de saída.

10. D. Direção especifica se a regra é aplicada ao tráfego de entrada ou de saída, o que torna a opção D a resposta certa. A opção A está incorreta; Ação é ou Permitir ou Negar. A opção B está incorreta; Alvo especifica o conjunto de instâncias ao qual a regra se aplica. A opção C está incorreta; Prioridade determina qual de todas as regras correspondentes é aplicada.
11. A. O 0.0.0.0/0 corresponde a todos os endereços IP, então a opção A está correta. A opção B representa um bloco de 16.777.214 endereços. A opção C representa um bloco de 1.048.574 endereços. A opção D representa um bloco de 65.534. Você pode experimentar com opções de bloco CIDR usando um calculador de CIDR como o que está disponível em www.subnet-calculator.com/cidr.php.
12. B. O produto com o qual você está trabalhando é compute e o recurso que você está criando é uma regra de firewall, então a opção B está correta. As opções A e C fazem referência a network em vez de compute. A opção D faz referência a rules em vez de firewall-rules.
13. B. O parâmetro correto é --network, o que torna a opção B correta. A opção A está incorreta; --subnet não é um parâmetro para o gcloud para criar um firewall. A opção C está incorreta; --destination não é um parâmetro válido. A opção D está incorreta; --source-ranges é para especificar fontes de tráfego de rede às quais a regra se aplica.
14. A. A regra na opção A usa o comando gcloud correto e especifica os parâmetros allow e direction. A opção B está incorreta porque faz referência a gcloud network em vez de gcloud compute. A opção C está incorreta porque não especifica a faixa de portas. A opção D está incorreta porque não especifica o protocolo ou a faixa de portas.
15. D. A opção D está correta porque é o maior número permitido na faixa de valores para prioridades. Quanto maior o número, menor a prioridade. Ter a menor prioridade garantirá que outras regras que correspondam serão aplicadas.
16. C. A opção de criar VPC está disponível na seção de Conectividade Híbrida, então a opção C está correta. Compute Engine, App Engine e IAM & Admin não têm recursos relacionados a VPNs.
17. B. A resposta correta é B; HA VPN é uma rede privada virtual que pode fornecer um SLA de disponibilidade de 99,99% e conectar redes locais ao Google Cloud. VPNs clássicas não fornecem alta disponibilidade, então a opção A está incorreta. A opção C, VPC compartilhada, é usada para tornar recursos em um projeto hospedeiro disponíveis para outros projetos. A opção D, emparelhamento de redes VPC, é usada para permitir um fluxo de tráfego entre VPCs, incluindo VPCs em diferentes organizações.

18. A. A opção A está correta porque o roteamento dinâmico global é usado para aprender todas as rotas em uma rede. A opção B está incorreta; o roteamento regional aprenderia apenas rotas em uma região. As opções C e D estão incorretas porque não são usadas para configurar opções de roteamento.
19. B. O comando correto é B, gcloud compute vpn-tunnels create. As opções A e C estão incorretas; gcloud network não é o início de um comando válido para criar túneis VPN. A opção D está incorreta; o termo create está na posição errada.
20. D. Ao usar gcloud para criar uma VPN, você precisa criar regras de encaminhamento, túneis e gateways, então todos os comandos gcloud listados seriam usados.

Capítulo 15: Rede na Nuvem: DNS, Balanceamento de Carga, Acesso Privado do Google e Endereçamento IP

1. **Registros DNS (B):** O registro A é essencial no DNS para vincular nomes de domínio a endereços IPv4, diferentemente do registro AAAA que liga a endereços IPv6, enquanto registros NS identificam servidores de nomes e SOA indicam a autoridade inicial de um domínio.
2. **DNSSEC (A):** Este protocolo assegura a autenticidade dos registros DNS, protegendo contra ataques como spoofing e envenenamento de cache, ao contrário de registros SOA e CNAME que não servem como medidas de segurança adicionais.
3. **Parâmetros TTL (A):** Definem quanto tempo um registro DNS pode ser armazenado em cache antes de ser necessário uma nova consulta ao servidor DNS, essencial para a eficiência e atualização dos registros DNS.
4. **Criação de Zonas Gerenciadas (B):** Utiliza-se o comando gcloud dns managed-zones create para criar uma zona DNS no Google Cloud, diferentemente do uso do gsutil que é direcionado ao gerenciamento do Cloud Storage.
5. **Visibilidade de Zonas DNS (B):** A visibilidade de uma zona DNS pode ser configurada como privada através do parâmetro de visibilidade, o que controla quem pode consultar os registros DNS dentro dessa zona.
6. **Balanceadores de Carga Globais (C):** Incluem o Balanceamento de Carga HTTP(S) Externo Global e clássico, SSL Proxy e TCP Proxy, fornecendo serviços de balanceamento de carga para aplicações acessíveis globalmente, em contraste com os平衡adores de carga regionais.
7. **Balanceamento de Carga Interno (D):** O Balanceamento de Carga HTTP(S) Interno distribui tráfego dentro de uma região na rede Premium, distinto dos balanceadores de carga globais que gerenciam tráfego internacional.
8. **Configuração de Privacidade no Console (A):** A configuração de privacidade no console do Google Cloud permite selecionar se o tráfego para as VMs vem da Internet ou apenas entre VMs, indicando a natureza pública ou privada do acesso.
9. **Configuração de Balanceadores de Carga de Proxy TCP (B):** Exige a configuração de componentes de frontend e backend, essencial para direcionar adequadamente o tráfego de entrada para os recursos corretos.
10. **Verificações de Saúde (B):** Monitoram a saúde das VMs associadas a balanceadores de carga, assegurando que o tráfego seja direcionado apenas para instâncias operacionais.
11. **Encaminhamento de Portas (B):** Ao configurar o frontend de um balanceador de carga, especifica-se quais portas devem ser encaminhadas, definindo como o tráfego de entrada é direcionado através do balanceador.

12. **Regras de Encaminhamento (A):** O comando gcloud compute forwarding-rules create é utilizado para estabelecer regras de encaminhamento no Google Cloud, essenciais para o direcionamento de tráfego de rede.
13. **Endereços Estáticos (C):** São atribuídos até serem explicitamente liberados, permitindo um endereçamento consistente para recursos como平衡adores de carga ou serviços expostos na internet.
14. **Endereços Efêmeros (A):** São adequados para recursos que não exigem acessibilidade externa consistente, como VMs de uso interno, podendo ser acessadas via SSH sem a necessidade de um endereço IP estático.
15. **Gerenciamento de Endereços IP (D):** Não é possível reduzir o número de endereços IP disponíveis através dos comandos fornecidos, refletindo a gestão fixa do espaço de endereçamento IP na configuração de sub-redes.
16. **Comprimento do Prefixo (B):** Define a porção da máscara de sub-rede de um endereço IP, influenciando diretamente o número de endereços IP disponíveis para dispositivos numa rede.
17. **Nível de Serviço de Rede Premium (C):** Roteia todo o tráfego pela infraestrutura global do Google, diferenciando-se do nível Standard que pode utilizar a Internet pública para roteamento de tráfego.
18. **Liberação de Endereços IP Efêmeros (B):** Parar e iniciar uma VM libera endereços IP efêmeros, sugerindo a utilização de endereços IP estáticos para manter a consistência de endereçamento entre reinicializações.
19. **Balanceamento de Carga Interno TCP/UDP (A):** É uma opção regional que suporta UDP, adequada para distribuição de tráfego interno, em contraste com平衡adores globais que focam em tráfego externo.
20. **Serviços de Rede no Console (B):** A seção de Serviços de Rede no console do Cloud abriga o console do Cloud DNS, permitindo a gestão de zonas DNS e registros para domínios hospedados no Google Cloud.

Capítulo 16: Implementando Aplicações com o Cloud Marketplace e o Cloud Foundation Toolkit

1. D. As categorias de soluções incluem todas as mencionadas, então a opção D está correta. Outras incluem Aplicações Kubernetes, API & Serviços e Bancos de Dados.
2. B. A resposta correta é B; o Cloud Foundation Toolkit fornece modelos e outras configurações para soluções comuns, como armazéns de dados, bem como templates para recursos específicos, como máquinas virtuais. A opção A, Cloud Deployment Manager, é um serviço para implementar soluções mas não é um conjunto de soluções exemplo. A opção C, Config Connector, é um complemento do Kubernetes para gerenciar recursos do Google Cloud a partir do Kubernetes. A opção D, Cloud Build, é um serviço do Google Cloud para construir containers.
3. A. Você inicia uma solução clicando no link "Launch On Compute Engine" na página de visão geral, então a opção A está correta. A opção B está incorreta; a página principal tem informações resumidas sobre os produtos. A opção C está incorreta; Serviços de Rede não está relacionado a este tópico. A opção D está incorreta porque a opção A é a resposta correta.
4. B. O Cloud Marketplace tem um conjunto de filtros predefinidos, incluindo filtragem por sistema operacional, então a opção B está correta. A opção A pode eventualmente levar à informação correta, mas não é eficiente. A opção D está incorreta porque é impraticável para uma tarefa tão simples.
5. B. Vários fornecedores podem oferecer configurações para as mesmas aplicações, então a opção B está correta. Isso dá aos usuários a oportunidade de escolher a mais adequada às suas necessidades. As opções A e C estão incorretas; isso é uma característica do Cloud Launcher. A opção D está incorreta porque a opção B é a resposta correta.
6. C. O Cloud Launcher exibirá opções de configuração apropriadas para a aplicação que você está implementando, então a opção C está correta. Por exemplo, ao implementar o WordPress, você terá a opção de implementar uma ferramenta de administração para PHP. A opção A está incorreta; isso é uma característica do Cloud Launcher. A opção B está incorreta; você não está necessariamente na página errada. A opção D está incorreta; isso é uma característica do Cloud Launcher.
7. D. Você pode alterar a configuração de qualquer um dos itens listados, então a opção D está correta. Você também pode especificar regras de firewall para permitir o tráfego HTTP e HTTPS ou mudar a zona em que a VM é executada.
8. B. Deployment Manager é o nome do serviço para criar recursos de aplicação usando um arquivo de configuração YAML, então a opção B está correta. A opção A está incorreta, embora você possa usar scripts com comandos gcloud para implementar recursos no Compute Engine. As opções C e D estão incorretas porque esses são nomes fictícios de produtos.

9. D. Arquivos de configuração são definidos na sintaxe YAML, então a opção D está correta.
10. B. Arquivos de configuração definem recursos e começam com a palavra resources, então a opção B está correta.
11. D. Todos os três—type, properties e name—são usados ao definir recursos em um arquivo de configuração do Cloud Deployment Manager, então a opção D está correta.
12. D. Todos os três podem ser configurados; especificamente, as chaves são deviceName, boot e autodelete. A opção D está correta.
13. A. A opção A é o comando correto. A opção B está incorreta; falta o termo compute. A opção C está incorreta; gsutil é o comando para trabalhar com o Cloud Storage. A opção D está incorreta porque os termos list e images estão na ordem errada.
14. D. O Google recomenda usar Python para templates complicados, então a opção D está correta. A opção A está incorreta porque Jinja2 é recomendado apenas para templates simples. As opções B e C estão incorretas; nenhuma das linguagens é suportada para templates.
15. A. A resposta correta é gcloud deployment-manager deployments create, então a opção A está correta. As opções B e D estão incorretas; o serviço não é chamado cloud-launcher no comando. A opção C está incorreta; launch não é um verbo válido para este comando.
16. C. A resposta correta é gcloud deployment-manager deployments describe, então a opção C está correta. As opções A e D estão incorretas; cloud-launcher não é o nome do serviço. A opção B está incorreta; list exibe um resumo breve de cada implementação. describe exibe uma descrição detalhada.
17. A. Você será capaz de configurar endereços IP, então a opção A está correta. Você não pode configurar faturamento ou controles de acesso no Deployment Manager, então as opções B e C estão incorretas. Você pode configurar o tipo de máquina, mas isso não está na seção Mais de Redes.
18. D. A resposta correta é a opção D porque gratuita, tarifa fixa por hora, taxas de uso e BYOL são todas opções de licença usadas no Cloud Marketplace.
19. B. Os tipos de licença tarifa fixa por hora e taxas de uso incluem o pagamento pela licença nas suas cobranças do Google Cloud, então a opção B está correta. O tipo de licença gratuita não incorre em cobranças. O tipo de licença BYOL exige que você trabalhe com o fornecedor do software para obter e pagar por uma licença. Não existe tal tipo de licença como chargeback, então a opção D está incorreta.
20. D. LAMP é a abreviação de Linux, Apache, MySQL e PHP. Todos estão incluídos ao instalar soluções LAMP, então a opção D está correta.

Capítulo 17: Configurando Acesso e Segurança

1. B. IAM significa gerenciamento de identidade e acesso, então a opção B está correta. A opção A está incorreta; o A não representa autorização, embora isso esteja relacionado. A opção C está incorreta; o A não representa auditoria, embora isso esteja relacionado. A opção D está incorreta. O IAM também trabalha com grupos, não apenas com indivíduos.
2. A. Membros e seus papéis são listados, então a opção A está correta. As opções B e C estão incorretas porque estão faltando a outra principal informação fornecida na listagem. A opção D está incorreta; permissões não são exibidas nessa página.
3. B. Papéis básicos foram criados antes do IAM e forneciam controles de acesso grosseiros, então a opção B está correta. A opção A está incorreta; eles são usados para controle de acesso. A opção C está incorreta; IAM é a forma mais nova de controle de acesso. A opção D está incorreta; eles fornecem funcionalidade de controle de acesso.
4. B. Papéis são usados para agrupar permissões que podem então ser atribuídas a identidades, então a opção B está correta. A opção A está incorreta; papéis não têm identidades, mas identidades podem ser concedidas papéis. A opção C está incorreta; papéis não usam listas de controle de acesso. A opção D está incorreta; papéis não incluem registros de auditoria. Os registros são coletados e gerenciados pelo Stackdriver Logging.
5. C. A resposta correta é gcloud projects get-iam-policy ace-exam-project, então a opção C está correta. A opção A está incorreta porque o recurso deve ser projects e não iam. A opção B está incorreta; list não fornece descrições detalhadas. A opção D está incorreta porque iam e list são referenciados incorretamente.
6. B. Novos membros podem ser usuários, indicados por seus endereços de email, ou grupos, então a opção B está correta. A opção A está incorreta; ela não inclui grupos. As opções C e D estão incorretas porque papéis não são adicionados lá.
7. D. Implementadores podem ler configurações e ajustes de aplicativos e escrever novas versões de aplicativos, então a opção D está correta. A opção A está incorreta porque está faltando a habilidade de ler configurações e ajustes. A opção B está incorreta porque está faltando escrever novas versões. A opção C está incorreta porque refere-se a escrever novas configurações.
8. B. Os passos corretos são navegar para IAM & Admin, selecionar Papéis e então marcar a caixa ao lado de um papel, então a opção B está correta. A opção A está incorreta; todos os papéis não são exibidos automaticamente. A opção C está incorreta; registros de auditoria não exibem permissões. A opção D está incorreta; não existe uma opção Papéis em Contas de Serviço.
9. D. Papéis predefinidos ajudam a implementar tanto o princípio do menor privilégio quanto a separação de deveres, então a opção D está correta. Papéis predefinidos não implementam defesa em profundidade por si só, mas podem ser usados com outros controles de segurança para implementar defesa em profundidade.

10. D. Os quatro estágios de lançamento disponíveis são alfa, beta, disponibilidade geral e desativado, então a opção D está correta.
11. B. A resposta correta, opção B, é `gcloud iam roles create`. A opção A está incorreta porque refere-se a `project` em vez de `iam`. A opção C está incorreta porque refere-se a `project` em vez de `iam`, e os termos `create` e `roles` estão fora de ordem. A opção D também está incorreta porque os termos `create` e `roles` estão fora de ordem.
12. B. Escopos são permissões concedidas a instâncias de VM, então a opção B está correta. Escopos em combinação com papéis IAM atribuídos a contas de serviço designadas para a instância de VM determinam quais operações a instância de VM pode realizar. As opções A e C estão incorretas; escopos não se aplicam a recursos de armazenamento. A opção D está incorreta; escopos não se aplicam a sub-redes.
13. C. Identificadores de escopo começam com `www.googleapis.com/auth` e são seguidos por um nome específico de escopo, como `devstorage.read_only` ou `logging.write`, então a opção C está correta. A opção A está incorreta; IDs de escopo não são gerados aleatoriamente. A opção B está incorreta; o nome de domínio não é `googleserviceaccounts`. A opção D está incorreta; escopos não são vinculados diretamente a projetos.
14. C. Tanto escopos quanto papéis IAM atribuídos a contas de serviço devem permitir uma operação para que ela tenha sucesso, então a opção C está correta. A opção A está incorreta; controles de acesso não afetam o fluxo de controle em aplicações a menos que explicitamente codificados para isso. A opção B está incorreta; a permissão mais permissiva não é usada. A opção D está incorreta; a operação não terá sucesso.
15. B. As opções para definir escopos são Permitir Acesso Padrão, Permitir Acesso Total e Definir Acesso Para Cada API, então a opção B está correta. A opção A está incorreta; está faltando Definir Acesso Para Cada API. A opção C está incorreta; está faltando Permitir Acesso Padrão. A opção D está incorreta; está faltando Permitir Acesso Total.
16. B. O comando correto é `gcloud compute instances set-service-account`, então a opção B está correta. A opção A está incorreta; não existe o verbo de comando `set-scopes`. A opção C está incorreta; o verbo de comando não é `set-scopes`. A opção D está incorreta; não existe o verbo de comando `define-scopes`.
17. A. Você pode atribuir uma conta de serviço ao criar uma VM usando o comando `create`. A opção B está incorreta; não existe o verbo de comando `create-service-account`. A opção C está incorreta; não existe o verbo de comando `define-service-account`. A opção D está incorreta; não existe o comando `instances-service-account`; além disso, `create` deve vir no final do comando.
18. C. O Cloud Logging coleta, armazena e exibe mensagens de log, então a opção C está correta. A opção A está incorreta; o Compute Engine não gerencia logs. A opção B está incorreta; o Cloud Storage não é usado para visualizar logs, embora arquivos de log possam ser armazenados lá. A opção D está incorreta; soluções personalizadas de log não são serviços do Google Cloud.

19. B. Logs podem ser filtrados por recurso, tipo de log, nível de log e período de tempo apenas, então a opção B está correta. As opções A, C e D estão incorretas porque cada uma delas está faltando pelo menos uma opção.
20. B. Este é um exemplo de atribuição do mínimo privilégio necessário para realizar uma tarefa, então a opção B está correta. A opção A está incorreta; defesa em profundidade combina múltiplos controles de segurança. A opção C está incorreta porque se trata de ter pessoas diferentes realizando tarefas sensíveis. A opção D está incorreta; varredura de vulnerabilidade é uma medida de segurança aplicada a aplicações que ajuda a revelar potenciais vulnerabilidades em uma aplicação que um atacante poderia explorar.

Capítulo 18: Monitoramento, Registro e Estimativa de Custos

1. **B.** O serviço de Monitoramento é utilizado para definir um limite em métricas e gerar alertas quando uma métrica excede o limite por um período de tempo especificado, portanto, a opção B está correta. A opção A está incorreta; o Registro é para coletar mensagens de log sobre eventos. A opção C está incorreta; o Cloud Trace é para rastreamento de aplicativos. A opção D está incorreta; o Debugger é um serviço descontinuado que era usado para depurar aplicativos e não estará mais disponível após maio de 2023.
2. **B.** Você instalaria o Agente de Operações na VM. O agente coleta dados e os envia para o Cloud Monitoring e Cloud Logging. Não existe algo como uma imagem de Operações na Nuvem. A opção de Monitorar Com Cloud Monitoring na página de configuração da VM não existe, e canais de notificação são definidos em uma política de alerta, não na VM.
3. **D.** O Cloud Monitoring pode monitorar recursos no Google Cloud, AWS e em data centers locais, portanto, a opção D está correta. As opções A até C estão incorretas porque não incluem duas outras opções corretas.
4. **A.** Um painel no Cloud Monitoring permitiria visualizar um conjunto de gráficos em um único lugar. Um sink do Cloud Logging é usado para rotear mensagens de log para um local de armazenamento, enquanto um Alerta de Monitoramento é para enviar notificações quando uma métrica excede ou fica abaixo de um limite.
5. **D.** Todas as três opções são canais de notificação válidos no Cloud Monitoring, incluindo o PagerDuty, uma ferramenta popular de DevOps.
6. **D.** A documentação é útil para documentar o propósito da política e fornecer orientação para resolver o problema. Onde uma política é armazenada é irrelevante para sua utilidade.
7. **A.** Fadiga de alerta é um estado causado por muitas notificações de alerta sendo enviadas para eventos que não requerem intervenção humana, criando o risco de que os engenheiros de DevOps comecem a prestar menos atenção às notificações.
8. **C.** O Cloud Logging armazena entradas de log no bucket Padrão por 30 dias.
9. **B.** A melhor opção é criar um bucket definido pelo usuário com uma política de retenção personalizada. Não há necessidade de escrever um script personalizado, levando mais tempo para desenvolver e manter do que usar a funcionalidade de exportação do Logging.
10. **D.** Todos os três, buckets do Cloud Storage, conjuntos de dados do BigQuery e tópicos do Cloud Pub/Sub, estão disponíveis como sinks para exportações de log.
11. **D.** Todas as opções listadas podem ser usadas para filtrar mensagens de log.
12. **B.** A resposta correta é paralisado. Não existe um nível padrão de log chamado status. Os status incluem Crítico, Erro, Aviso, Informação e Depuração.

13. A. Você pode expandir o campo metadataRequest na estrutura JSON da mensagem no Log Explorer. O Metric Explorer é usado com métricas do Cloud Monitoring, não com mensagens de log.
14. C. O Cloud Trace é um aplicativo de rastreamento distribuído que fornece detalhes sobre quanto tempo diferentes partes do código são executadas. O Monitoramento é usado para notificar os engenheiros de DevOps quando os recursos não estão funcionando conforme o esperado.
15. C. Trace é um aplicativo de rastreamento distribuído que fornece detalhes sobre a duração da execução de diferentes partes do código. O Monitoramento é usado para observar métricas sobre serviços e alertar sobre condições indesejadas.
16. B. O Painel de Status do Google Cloud em <https://status.cloud.google.com> tem informações sobre o status dos serviços do Google Cloud.
17. B. Tanto o Compute Engine quanto o Kubernetes Engine exigirão detalhes sobre as configurações das VMs. BigQuery e Cloud Pub/Sub são serviços sem servidor.
18. C. A especificação de consultas no BigQuery é baseada na quantidade de dados escaneados, então a opção C está correta. A opção A está incorreta; a quantidade de armazenamento de dados é especificada na seção de Precificação de Armazenamento. A opção B está incorreta; a precificação de consultas não é baseada no volume de dados retornados. A opção D está incorreta porque o número de partições não é um fator na precificação do BigQuery.
19. B. Alguns sistemas operacionais, como o Microsoft Windows Server, requerem uma licença, então a opção B está correta. O Google às vezes tem acordos com fornecedores para cobrar taxas pelo uso de software proprietário. A opção A está incorreta; não há uma tarifa fixa para sistemas operacionais. A opção C está incorreta; a informação é às vezes necessária para calcular as cobranças. A opção D está incorreta porque, se você trazer sua própria licença, não haverá cobrança adicional de licença.
20. D. A opção D está correta. O bucket Obrigatório é usado para armazenar atividades administrativas, eventos do sistema e transparência de acesso. A opção A está incorreta; mensagens do sistema operacional não são direcionadas para o bucket Obrigatório. As opções B e C estão incorretas porque incluem apenas alguns dos tipos de logs direcionados para o bucket Obrigatório.