PassQueen
a.k.a dumpsqueen.com

Google Certified Professional - Cloud Architect (GCP)

Google Professional-Cloud-Architect

Version Demo

Total Demo Questions: 40

Total Premium Questions: 401

Buy Premium PDF

https://passqueen.com

support@passqueen.com

Topic Break Down

Topic	No. of Questions
Topic 1, Designing and planning a cloud solution architecture	45
Topic 2, Managing and provisioning the solution infrastructure	46
Topic 3, Designing for security and compliance	67
Topic 4, Analyzing and optimizing technical and business processes	20
Topic 5, Managing implementation	24
Topic 6, Ensuring solution and operations reliability	50
Topic 7, Mix Questions	1
Topic 8, Case Study 1	50
Topic 9, Case Study 2	59
Topic 10, Case Study 3	17
Topic 11, Case Study 4	10
Topic 12, Case Study 5	12
Total	401

QUESTION NO: 1

For this question, refer to the TerramEarth case study.

The TerramEarth development team wants to create an API to meet the company ' s business requirements. You want the development team to focus their development effort on business value versus creating a custom framework. Which method should they use?

A. Use Google App Engine with Google Cloud Endpoints. Focus on an API for dealers and partners.

B. Use Google App Engine with a JAX-RS Jersey Java-based framework. Focus on an API for the public.

C. Use Google App Engine with the Swagger (open API Specification) framework. Focus on an API for the public.

D. Use Google Container Engine with a Django Python container. Focus on an API for the public.

E. Use Google Container Engine with a Tomcat container with the Swagger (Open API Specification) framework. Focus on an API for dealers and partners.

ANSWER: A

Explanation:

Use Google App Engine with Google Cloud Endpoints. Focus on an API for dealers and partners. This approach fits TerramEarth's need to expose services to its dealer network and external business partners, including seed and fertilizer suppliers, while minimizing undifferentiated platform work. Cloud Endpoints provides managed API capabilities such as API definition support, authentication, authorization integration, traffic management, monitoring, logging, and client-library generation. As a result, the team can concentrate on implementing equipment, inventory, dealer, and partnership business capabilities rather than building and operating these cross-cutting API framework concerns themselves.

App Engine is also appropriate for this workload because it provides a managed application platform that handles infrastructure provisioning, scaling, and operational overhead. Combining it with Cloud Endpoints gives the development team a practical managed path for publishing secure, observable APIs. The intended consumers should be dealers and partners because the stated business objectives emphasize helping dealers better understand customer equipment use and enabling joint offerings with other companies. Google documents Cloud Endpoints as a platform for developing, deploying, protecting, and monitoring APIs, with support for OpenAPI-based service configuration. See Cloud Endpoints overview and App Engine documentation.

QUESTION NO: 2

You are developing your microservices application on Google Kubernetes Engine. During testing, you want to validate the behavior of your application in case a specific microservice should suddenly crash. What should you do?

A. Add a taint to one of the nodes of the Kubernetes cluster. For the specific microservice, configure a pod anti-affinity label that has the name of the tainted node as a value.

B. Use Istio's fault injection on the particular microservice whose faulty behavior you want to simulate.

C. Destroy one of the nodes of the Kubernetes cluster to observe the behavior.

D. Configure Istio's traffic management features to steer the traffic away from a crashing microservice.

ANSWER: B

Explanation:

Use Istio's fault injection on the particular microservice whose faulty behavior you want to simulate. Fault injection is designed for resilience testing in a service mesh: it deliberately introduces controlled failures into requests directed at a selected service, such as HTTP aborts or latency delays. This lets the team observe whether callers handle errors correctly, whether retries, timeouts, circuit breaking, fallback behavior, and alerting operate as intended, and whether the remainder of the application remains available when that service is effectively failing.

Because the fault rule can be scoped to a particular workload, route, and percentage of traffic, testing can be targeted and repeatable without disrupting unrelated services or requiring destructive changes to the Kubernetes cluster. For a sudden-failure scenario, an HTTP abort fault can emulate the failed responses clients would receive when the service is unavailable. This is a standard approach to validating microservice failure handling before production. Google Cloud Service Mesh documentation describes fault injection as a way to test application resilience by injecting delays and aborts into requests, while Istio documents the same mechanisms and their traffic-routing configuration. See Google Cloud Service Mesh fault injection documentation and Istio fault injection documentation.

QUESTION NO: 3

TerramEarth has about 1 petabyte (PB) of vehicle testing data in a private data center. You want to move the data to Cloud Storage for your machine learning team. Currently, a 1-Gbps interconnect link is available for you. The machine learning team wants to start using the data in a month. What should you do?

A. Request Transfer Appliances from Google Cloud, export the data to appliances, and return the appliances to Google Cloud.

B. Configure the Storage Transfer service from Google Cloud to send the data from your data center to Cloud Storage

C. Make sure there are no other users consuming the 1 Gbps link, and use multi-thread transfer to upload the data to Cloud Storage.

D. Export files to an encrypted USB device, send the device to Google Cloud, and request an import of the data to Cloud Storage

ANSWER: A

Explanation:

Request Transfer Appliances from Google Cloud, export the data to appliances, and return the appliances to Google Cloud. This is the appropriate offline transfer approach for moving approximately 1 PB within the required timeline. Even under ideal conditions, transferring 1 PB through a dedicated 1 Gbps connection would take roughly 93 days; real-world overhead, protocol behavior, and link utilization would make the elapsed time longer. Therefore, the available network link cannot reliably meet a one-month deadline.

Transfer Appliance is designed for large-volume migrations when network transfer is impractical or would take too long. Google ships secure, high-capacity appliances to the customer site, where the data is copied locally. The appliances are then returned to Google, which transfers the data into the designated Cloud Storage destination. This removes the 1 Gbps link as the limiting factor and lets the team begin using the migrated vehicle testing data much sooner. Transfer Appliance also supports encryption and chain-of-custody processes intended for secure physical data migration. See the Transfer Appliance overview and Google Cloud's Transfer Appliance documentation.

QUESTION NO: 4

Your company has a Google Cloud project that uses BigQuery for data warehousing. They have a VPN tunnel between the on-premises environment and Google Cloud that is configured with Cloud VPN.

The security team wants to avoid data exfiltration by malicious insiders, compromised code, and accidental oversharing. What should they do?

A. Configure Private Google Access for on-premises only.

B. Perform the following tasks:
1. Create a service account.
2. Give the BigQuery JobUser role and Storage Reader role to the service account.
3. Remove all other IAM access from the project.

C. Configure VPC Service Controls and configure Private Google Access.

D. Configure Private Google Access.

ANSWER: C

Explanation:

Configure VPC Service Controls and configure Private Google Access. VPC Service Controls creates a security perimeter around supported Google Cloud services, including BigQuery, and provides an additional layer of defense beyond IAM. The perimeter can restrict access to protected BigQuery resources to requests originating from authorized networks, projects, and access levels. This directly helps limit unintended movement of sensitive data caused by compromised workloads, malicious users with otherwise valid credentials, or incorrectly shared resources.

For an on-premises environment connected through Cloud VPN, Private Google Access enables systems using private IP addresses to reach Google APIs and services without requiring external IP addresses. In a VPC Service Controls design, private access should be configured to use the restricted Google APIs endpoint where appropriate. This ensures that API traffic to protected services remains on Google's network and is subject to the service perimeter, rather than allowing broad access through public API endpoints. Together, the service perimeter and private API connectivity provide the intended defense-in-depth controls for BigQuery data-exfiltration protection. See the VPC Service Controls overview and Google Cloud guidance for Private Google Access.

QUESTION NO: 5

Your company runs a customer-facing application on Google Kubernetes Engine (GKE). The application must remain available during a single-zone failure and during planned node upgrades. You want to use managed Google Cloud capabilities and avoid maintaining spare virtual machines manually.

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

A. Create a regional GKE cluster.

B. Configure a PodDisruptionBudget for each replicated application.

C. Deploy all application replicas to one node by using pod affinity.

D. Use a single zonal node pool and manually restart failed nodes.

E. Disable GKE node auto-upgrades permanently.

ANSWER: A B

Explanation:

Create a regional GKE cluster because its control plane and default node pools can be distributed across multiple zones in the selected region. This reduces the impact of a single-zone failure compared with a zonal cluster. For a customer-facing workload, replicas should be spread across the available zones so that healthy Pods remain available when one zone is unavailable.

Configure a PodDisruptionBudget for each replicated application. A PodDisruptionBudget limits the number of replicas that can be voluntarily disrupted at one time, such as during a node upgrade or node-pool maintenance operation. This helps ensure that sufficient healthy application replicas remain to serve traffic while GKE performs managed maintenance. See Regional GKE clusters and PodDisruptionBudgets.

QUESTION NO: 6

You have a Python web application with many dependencies that requires 0.1 CPU cores and 128 MB of memory to operate in production. You want to monitor and maximize machine utilization. You also want to reliably deploy new versions of the application. Which set of steps should you take?

A. Perform the following:
1. Create a managed instance group with f1-micro type machines.
2. Use a startup script to clone the repository, check out the production branch, install the dependencies, and start the Python app.
3. Restart the instances to automatically deploy new production releases.

B. Perform the following:
1. Create a managed instance group with n1-standard-1 type machines.
2. Build a Compute Engine image from the production branch that contains all of the dependencies and automatically starts the Python app.
3. Rebuild the Compute Engine image, and update the instance template to deploy new production releases.

C. Perform the following:
1. Create a Google Kubernetes Engine (GKE) cluster with n1-standard-1 type machines.
2. Build a Docker image from the production branch with all of the dependencies, and tag it with the version number.
3. Create a Kubernetes Deployment with the imagePullPolicy set to 'IfNotPresent' in the staging namespace, and then deploy the same version-tagged image to the production namespace after testing.

D. Perform the following:
1. Create a GKE cluster with n1-standard-4 type machines.
2. Build a Docker image from the master branch with all of the dependencies, and tag it with 'latest'.
3. Create a Kubernetes Deployment in the default namespace with the imagePullPolicy set to 'Always'. Restart the pods to automatically deploy new production releases.

ANSWER: C

Explanation:

Creating a GKE cluster with n1-standard-1 type machines, building a version-tagged Docker image, and deploying it through Kubernetes Deployments is the best fit for a small application that needs efficient resource use and dependable releases. Kubernetes schedules multiple Pods onto each node according to their declared CPU and memory requests. With requests set near the application's 0.1 vCPU and 128 MB requirements, many replicas or other workloads can share a node rather than leaving most of a VM unused. GKE also exposes workload and node metrics, enabling capacity monitoring and scaling decisions.

A container image packages the Python runtime and all application dependencies into an immutable, repeatable deployment artifact. Tagging that image with a specific version ensures that the tested artifact is identifiable and can be deployed consistently. A Kubernetes Deployment manages the desired replica count and supports controlled rolling updates, health-based replacement, rollout status checks, and rollback to an earlier revision when needed. Testing the version-tagged workload in staging before deploying the same artifact to production provides a reliable release path. See the GKE resource management documentation and the Kubernetes Deployment documentation.

QUESTION NO: 7

You are using Cloud SQL as the database backend for a large CRM deployment. You want to scale as usage increases and ensure that you don't run out of storage, maintain 75% CPU usage cores, and keep replication lag below 60 seconds. What are the correct steps to meet your requirements?

A. 1. Enable automatic storage increase for the instance.
2. Create a Stackdriver alert when CPU usage exceeds 75%, and change the instance type to reduce CPU usage.
3. Create a Stackdriver alert for replication lag, and shard the database to reduce replication time.

B. 1. Enable automatic storage increase for the instance.
2. Change the instance type to a 32-core machine type to keep CPU usage below 75%.
3. Create a Stackdriver alert for replication lag, and deploy memcache to reduce load on the master.

C. 1. Create a Stackdriver alert when storage exceeds 75%, and increase the available storage on the instance to create more space.
2. Deploy memcached to reduce CPU load.
3. Change the instance type to a 32-core machine type to reduce replication lag.

D. 1. Create a Stackdriver alert when storage exceeds 75%, and increase the available storage on the instance to create more space.
2. Deploy memcached to reduce CPU load.
3. Create a Stackdriver alert for replication lag, and change the instance type to a 32-core machine type to reduce replication lag.

ANSWER: A

Explanation:

Enabling automatic storage increase is the appropriate first step because Cloud SQL can automatically add storage capacity as the instance approaches its configured storage limit. This protects the CRM workload from an outage caused by exhausted database storage while allowing growth without requiring a manual resize at a fixed utilization threshold.

CPU utilization and replication lag are workload-dependent operational signals, so they should be monitored with Cloud Monitoring alerts. An alert at the defined CPU threshold provides a measured trigger to resize the Cloud SQL instance to a larger machine type when sustained demand requires additional compute capacity. Similarly, an alert for replication lag makes it possible to act before the required 60-second service target is breached. Sharding distributes data and write activity across separate database partitions, reducing the amount of work handled by an individual primary/replication stream and providing a horizontal scaling approach as CRM usage grows. This combines automated capacity growth for storage with observable, threshold-driven scaling for compute and replication performance. See Cloud SQL automatic storage increase and Cloud SQL monitoring documentation.

QUESTION NO: 8

You are helping the QA team to roll out a new load-testing tool to test the scalability of your primary cloud services that run on Google Compute Engine with Cloud Bigtable.

Which three requirements should they include? (Choose three.)

A. Ensure that the load tests validate the performance of Cloud Bigtable

B. Create a separate Google Cloud project to use for the load-testing environment

C. Schedule the load-testing tool to regularly run against the production environment

D. Ensure all third-party systems your services use is capable of handling high load

E. Instrument the production services to record every transaction for replay by the load-testing tool

F. Instrument the load-testing tool and the target services with detailed logging and metrics collection

ANSWER: A B F

Explanation:

Load testing must measure the behavior of the complete primary service path, including Cloud Bigtable. Cloud Bigtable performance depends on workload characteristics such as read/write patterns, row-key design, throughput, and hotspotting; therefore, a test that assesses only the Compute Engine tier would not establish whether the overall service can scale. Google provides guidance and tools for benchmarking Bigtable workloads in its Cloud Bigtable performance documentation.

A separate Google Cloud project is an appropriate boundary for a load-testing environment. It enables isolation of test resources, IAM permissions, quotas, billing attribution, and operational changes from production. The environment should be sufficiently representative to produce meaningful results, while avoiding unnecessary production impact.

Detailed metrics and logging in both the load generator and the services under test are essential to interpret results. The test needs to correlate offered load, latency, error rates, resource utilization, and Bigtable-related signals to identify saturation points and validate scaling behavior. Google Cloud observability guidance describes using telemetry to understand application and infrastructure health in the Operational Excellence Framework.

QUESTION NO: 9

You need to implement a network ingress for a new game that meets the defined business and technical requirements. Mountkirk Games wants each regional game instance to be located in multiple Google Cloud regions. What should you do?

A. Configure a global load balancer connected to a managed instance group running Compute Engine instances.

B. Configure kubemci with a global load balancer and Google Kubernetes Engine.

C. Configure a global load balancer with Google Kubernetes Engine.

D. Configure Ingress for Anthos with a global load balancer and Google Kubernetes Engine.

ANSWER: A

Explanation:

Configure a global load balancer connected to a managed instance group running Compute Engine instances. A global external Application Load Balancer provides a single anycast IP address for players while directing requests to healthy backends across multiple Google Cloud regions. The load balancer uses Google's global network and health checks to select an appropriate regional backend, helping provide low-latency access and continuity when a backend instance or an entire regional backend becomes unhealthy.

Managed instance groups are the appropriate Compute Engine backend construct because they maintain a homogeneous fleet of game-serving virtual machines. A managed instance group can automatically recreate failed VMs, and it can be configured with autoscaling so each regional fleet grows or shrinks with player demand. Regional managed instance groups also distribute their instances across multiple zones within a region, adding zonal resilience; deploying equivalent groups in multiple regions and attaching them to the same global load-balancing backend service provides the requested multi-region game ingress architecture. This design combines global traffic steering with managed, scalable Compute Engine capacity. See Google Cloud external Application Load Balancer documentation and Google Cloud managed instance groups documentation.

QUESTION NO: 10

For this question, refer to the Dress4Win case study. You want to ensure that your on-premises architecture meets business requirements before you migrate your solution.

What change in the on-premises architecture should you make?

A. Replace RabbitMQ with Google Pub/Sub.

B. Downgrade MySQL to v5.7, which is supported by Cloud SQL for MySQL.

C. Resize compute resources to match predefined Compute Engine machine types.

D. Containerize the micro-services and host them in Google Kubernetes Engine.

ANSWER: C

Explanation:

Resize compute resources to match predefined Compute Engine machine types. is the appropriate change because migration planning should validate that each workload's CPU and memory requirements can be mapped to practical target virtual machine configurations before moving production traffic. Aligning the on-premises resource profile with Compute Engine machine types makes capacity planning, cost estimation, performance testing, and operational sizing more predictable. It also helps expose workloads that are overprovisioned, unusually sized, or dependent on hardware characteristics that need a deliberate target design. Compute Engine provides predefined machine types with defined vCPU and memory combinations, allowing the organization to select and test a repeatable configuration that satisfies its business performance requirements. Where a predefined size is not a close fit, Compute Engine also supports custom machine types, but right-sizing remains an essential pre-migration activity. Completing this assessment while the application is still on premises reduces the risk of discovering resource constraints only after cutover and provides a sound baseline for subsequent scaling design. See the Compute Engine machine resource documentation and Google Cloud's migration assessment guidance.

QUESTION NO: 11

For this question, refer to the JencoMart case study

A few days after JencoMart migrates the user credentials database to Google Cloud Platform and shuts down the old server, the new database server stops responding to SSH connections. It is still serving database requests to the application servers correctly. What three steps should you take to diagnose the problem? Choose 3 answers

A. Delete the virtual machine (VM) and disks and create a new one.

B. Delete the instance, attach the disk to a new VM, and investigate.

C. Take a snapshot of the disk and connect to a new machine to investigate.

D. Check inbound firewall rules for the network the machine is connected to.

E. Connect the machine to another network with very simple firewall rules and investigate.

F. Print the Serial Console output for the instance for troubleshooting, activate the interactive console, and investigate.

ANSWER: C D F

Explanation:

Because the database continues serving application traffic, the instance and its workload are still running; the investigation should focus on the SSH access path and use non-disruptive recovery mechanisms. Checking inbound firewall rules for the network the machine is connected to verifies that TCP port 22 is allowed from the administrator's source address and that relevant network tags or service-account-targeted rules apply to the instance. Firewall configuration is a primary component of Compute Engine SSH connectivity troubleshooting.

Printing the Serial Console output for the instance for troubleshooting, activating the interactive console, and investigating provides out-of-band access independent of the network and SSH daemon. This can expose boot errors, OS-level network configuration problems, SSH daemon failures, or account and key configuration issues without stopping a healthy database workload.

Taking a snapshot of the disk and connecting to a new machine to investigate preserves the original persistent disk state while allowing offline inspection from a separate VM. Administrators can examine system and SSH logs, configuration files, authorized keys, and disk-space or filesystem conditions without risking changes to the production instance. Google documents firewall checks, serial-console access, and disk-based recovery as useful SSH troubleshooting approaches. See Troubleshoot SSH and Create disk snapshots.

QUESTION NO: 12

Your agricultural division is experimenting with fully autonomous vehicles. You want your architecture to promote strong security during vehicle operation. Which two architectures should you consider? (Choose two.)

A. Treat every micro service call between modules on the vehicle as untrusted.

B. Require IPv6 for connectivity to ensure a secure address space.

C. Use a trusted platform module (TPM) and verify firmware and binaries on boot.

D. Use a functional programming language to isolate code execution cycles.

E. Use multiple connectivity subsystems for redundancy.

F. Enclose the vehicle's drive electronics in a Faraday cage to isolate chips.

ANSWER: A C

Explanation:

Treat every micro service call between modules on the vehicle as untrusted. applies a zero-trust architecture to the vehicle's internal systems. Vehicle modules, services, and workloads should not receive implicit trust merely because they run on the same device or internal network. Each request should be authenticated and authorized, with strong workload identities, encrypted service-to-service communication, and least-privilege access. This limits the impact of a compromised component and helps prevent lateral movement to safety-critical control functions. Google's zero-trust guidance emphasizes continuously verifying access rather than relying on network location or perimeter trust. See Google Cloud's zero-trust security model.

Use a trusted platform module (TPM) and verify firmware and binaries on boot. establishes a hardware-rooted chain of trust. A TPM can securely store cryptographic keys and measurements of boot components, while verified or measured boot checks that firmware, the bootloader, operating system, and deployed binaries have not been altered before allowing vehicle software to run. This is particularly important for autonomous vehicles, where compromise of low-level firmware or control software can undermine all higher-level controls. Secure boot and integrity verification provide a defensible starting state for every operating cycle. Google documents the importance of verified boot and a hardware root of trust in Android Verified Boot.

QUESTION NO: 13

Your company runs several applications on Google Kubernetes Engine (GKE). The applications need to retrieve database passwords and API tokens at runtime. You want to avoid storing long-lived credentials in container images, Kubernetes manifests, or Kubernetes Secrets.

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

A. Store the passwords and API tokens in Secret Manager.

B. Configure Workload Identity Federation for GKE for the application workloads.

C. Embed a service account key in each container image so the application can access the secrets.

D. Store the credentials in a ConfigMap and mount the ConfigMap into each Pod.

E. Grant the Compute Admin role to the default Compute Engine service account.

ANSWER: A B

Explanation:

Store the credentials in Secret Manager and grant the minimum required Secret Manager access to a dedicated Google Cloud service account. Secret Manager is designed for centrally storing, managing, and auditing access to sensitive values such as passwords and API tokens.

Configure Workload Identity Federation for GKE to map the Kubernetes service account used by the workload to the Google Cloud service account. The workload can then obtain short-lived credentials to call Google Cloud APIs without distributing service account key files to containers. See Secret Manager overview and Workload Identity Federation for GKE.

QUESTION NO: 14

Your company stores sensitive research data in Cloud Storage and analyzes it with BigQuery. The security team must reduce the risk that users with valid credentials can copy this data to unauthorized Google Cloud projects or external services. The analytics team must continue to use BigQuery and Cloud Storage.

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

A. Create a VPC Service Controls service perimeter that includes the Cloud Storage and BigQuery projects.

B. Configure perimeter ingress and egress rules that allow only approved access paths.

C. Grant the Storage Admin and BigQuery Admin roles to all analysts.

D. Disable Cloud Audit Logs for the protected projects.

E. Copy all sensitive data into each analyst's personal project.

ANSWER: A B

Explanation:

Create a VPC Service Controls service perimeter that includes the projects containing the sensitive Cloud Storage and BigQuery resources. VPC Service Controls provides an additional security boundary around supported Google Cloud services and can restrict access to protected resources from outside the perimeter. This helps mitigate data exfiltration risks that IAM alone does not address.

Configure ingress and egress rules for the perimeter that allow only the required identities, projects, and services. These rules let the organization preserve approved analytics workflows while preventing unrestricted transfers to unapproved projects or external destinations. The rules should be narrowly scoped and validated against the required operational paths. See VPC Service Controls overview and VPC Service Controls ingress and egress rules.

QUESTION NO: 15

Your marketing department wants to send out a promotional email campaign. The development team wants to minimize direct operation management. They project a wide range of possible customer responses, from 100 to 500,000 click-through per day. The link leads to a simple website that explains the promotion and collects user information and preferences. Which infrastructure should you recommend? (Choose two.)

A. Use Google App Engine to serve the website and Google Cloud Datastore to store user data.

B. Use a Google Container Engine cluster to serve the website and store data to persistent disk.

C. Use a managed instance group to serve the website and Google Cloud Bigtable to store user data.

D. Use a single Compute Engine virtual machine (VM) to host a web server, backend by Google Cloud SQL.

ANSWER: A C

Explanation:

Use Google App Engine to serve the website and Google Cloud Datastore to store user data. is appropriate because App Engine is a fully managed application platform that automatically scales web applications in response to demand. This minimizes server provisioning, patching, capacity management, and load-balancer administration while handling potentially dramatic campaign-driven traffic variation. Cloud Datastore, now provided through Firestore in Datastore mode, is a managed, horizontally scalable NoSQL database well suited to storing user-submitted profile and preference records without managing database servers.

Use a managed instance group to serve the website and Google Cloud Bigtable to store user data. is also a viable scalable architecture. A managed instance group can automatically add or remove Compute Engine instances according to load and can be used with Google Cloud load balancing for a resilient web tier. Cloud Bigtable is a fully managed, high-throughput, low-latency wide-column database that can scale to accommodate very large request volumes and data growth. This combination supports the projected peak traffic while using managed scaling and managed data services rather than requiring the team to operate individual web-server instances or database infrastructure. See App Engine automatic scaling and Cloud Bigtable overview.

QUESTION NO: 16

Your development team has created a mobile game app. You want to test the new mobile app on Android and iOS devices with a variety of configurations. You need to ensure that testing is efficient and cost-effective. What should you do?

A. Upload your mobile app to the Firebase Test Lab, and test the mobile app on Android and iOS devices.

B. Create Android and iOS VMs on Google Cloud, install the mobile app on the VMs, and test the mobile app.

C. Create Android and iOS containers on Google Kubernetes Engine (GKE), install the mobile app on the containers, and test the mobile app.

D. Upload your mobile app with different configurations to Firebase Hosting and test each configuration.

ANSWER: A

Explanation:

Upload your mobile app to the Firebase Test Lab, and test the mobile app on Android and iOS devices. Firebase Test Lab is a managed cloud testing service designed specifically for running mobile application tests across a broad device matrix without the operational cost of buying, provisioning, maintaining, and continually updating a physical device lab. It supports testing Android apps on both virtual and physical devices, and supports iOS testing on physical devices. This lets the team select device models, OS versions, locales, orientations, and other relevant configurations that represent its users.

The team can upload build artifacts and run automated tests, such as instrumentation tests and game-loop tests for Android games, at scale. Test Lab collects actionable artifacts including test results, logs, screenshots, and videos, enabling developers to identify device-specific failures efficiently. Its managed, on-demand model is especially suitable when the required test matrix varies between releases, because capacity is available without persistent infrastructure expense. Firebase also provides quotas and billing controls so the organization can balance test coverage and cost. See the Firebase Test Lab documentation and the iOS testing overview.

QUESTION NO: 17

To reduce costs, the Director of Engineering has required all developers to move their development infrastructure resources from on-premises virtual machines (VMs) to Google Cloud Platform. These resources go through multiple start/stop events during the day and require state to persist. You have been asked to design the process of running a development environment in Google Cloud while providing cost visibility to the finance department. Which two steps should you take? Choose 2 answers

A. Use the --no-auto-delete flag on all persistent disks and stop the VM.

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

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

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

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

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

ANSWER: A D

Explanation:

“Use the --no-auto-delete flag on all persistent disks and stop the VM.” preserves the development environment’s state while avoiding compute charges when developers are not actively using their VMs. Stopping a Compute Engine VM retains its attached Persistent Disk data, and configuring disks not to auto-delete also protects those disks if the VM is later deleted as part of lifecycle management. This supports repeated start/stop usage without rebuilding the environment each time. Persistent Disk storage charges continue while a VM is stopped, but the VM’s vCPU and memory charges do not, making this an appropriate cost-control approach for intermittent development workloads.

“Use Google BigQuery billing export and labels to associate cost to groups.” provides the finance department with detailed, queryable cost visibility. Billing export sends detailed Google Cloud billing usage and cost data to BigQuery. Consistently applied resource labels, such as team, project, cost center, or environment, are included with eligible usage records and enable finance teams to allocate, filter, aggregate, and report costs by organizational group. Together, labels and BigQuery billing export establish a scalable chargeback or showback process across development resources.

See Compute Engine disk auto-delete behavior and Export Cloud Billing data to BigQuery.

QUESTION NO: 18

The JencoMart security team requires that all Google Cloud Platform infrastructure is deployed using a least privilege model with separation of duties for administration between production and development resources.

What Google domain and project structure should you recommend?

A. Create two G Suite accounts to manage users: one for development/test/staging and one for production. Each account should contain one project for every application

B. Create two G Suite accounts to manage users: one with a single project for all development applications and one with a single project for all production applications

C. Create a single G Suite account to manage users with each stage of each application in its own project

D. Create a single G Suite account to manage users with one project for the development/test/staging environment and one project for the production environment

ANSWER: C

Explanation:

Create a single G Suite account to manage users with each stage of each application in its own project. A single Google Workspace (formerly G Suite) domain establishes one central identity and organizational boundary, allowing consistent lifecycle management for users, groups, authentication policies, and organization-level governance. Separating every application stage into its own project creates strong resource and IAM policy boundaries: production resources are isolated from development, test, and staging resources, and one application's environment is also isolated from another application's environment.

This structure supports least privilege because administrators, developers, and automation identities can receive narrowly scoped roles only in the specific project and environment where their work is required. It also supports separation of duties by allowing distinct groups, such as production operators and development administrators, to be granted access to their respective projects without inheriting unnecessary permissions over other environments. Projects are a fundamental Google Cloud resource hierarchy boundary and are designed to contain resources, identities, permissions, quotas, and billing configuration. This granular project-per-application-stage model limits administrative blast radius while retaining centralized identity management. See Google Cloud resource hierarchy and IAM security best practices.

QUESTION NO: 19

Your organization deploys containerized applications to GKE. The security team requires that only images produced by the approved CI/CD pipeline can be deployed, and that images are evaluated for known vulnerabilities before release.

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

A. Enable vulnerability scanning for container images stored in Artifact Registry.

B. Configure Binary Authorization for GKE to require an attestation from the approved CI/CD pipeline.

C. Grant every developer the GKE Cluster Admin role so they can approve their own deployments.

D. Allow GKE nodes to pull container images directly from public container registries.

E. Store container images in a Cloud Storage bucket instead of Artifact Registry.

ANSWER: A B

Explanation:

Enable vulnerability scanning for images stored in Artifact Registry. Artifact Analysis can identify known vulnerabilities in container images and provide findings that the delivery process can evaluate before a release is approved.

Configure Binary Authorization for GKE and require an attestation from the approved CI/CD pipeline. Binary Authorization enforces a policy at deployment time, preventing images that do not satisfy the configured attestation requirements from being deployed to GKE. Together, these controls provide vulnerability visibility and an enforceable admission policy for approved artifacts. See Artifact Analysis container scanning and Binary Authorization overview.

QUESTION NO: 20

Your company wants to try out the cloud with low risk. They want to archive approximately 100 TB of their log data to the cloud and test the analytics features available to them there, while also retaining that data as a long-term disaster recovery backup.

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

A. Load logs into Google BigQuery

B. Load logs into Google Cloud SQL

C. Import logs into Google Stackdriver

D. Insert logs into Google Cloud Bigtable

E. Upload log files into Google Cloud Storage

ANSWER: A E

Explanation:

Uploading the log files into Google Cloud Storage provides durable, scalable object storage for a 100 TB archive and disaster-recovery copy. Cloud Storage is designed to store arbitrary file data at very large scale, and lifecycle management can transition objects to lower-cost storage classes as the logs age. This supports a low-risk initial migration because the original files can be retained in their native form, independently of any subsequent analytics processing.

Loading logs into Google BigQuery enables the company to test cloud analytics using a serverless data warehouse that can run SQL queries over very large datasets without managing database infrastructure. Log files stored in Cloud Storage can be loaded into BigQuery tables for structured analysis, while Cloud Storage remains the long-term archive and recovery source. This separation is useful operationally: BigQuery supplies interactive analysis and can be populated or refreshed as needed, whereas Cloud Storage holds the durable source files for retention. Google documents Cloud Storage as object storage for data lakes, backup, and archive workloads, and documents BigQuery loading from Cloud Storage for batch ingestion and analysis. See Cloud Storage documentation and Load data from Cloud Storage into BigQuery.

QUESTION NO: 21

You want to establish a Compute Engine application in a single VPC across two regions. The application must communicate over VPN to an on-premises network. How should you deploy the VPN?

A. Use VPC Network Peering between the VPC and the on-premises network.

B. Expose the VPC to the on-premises network using IAM and VPC Sharing.

C. Create a global Cloud VPN Gateway with VPN tunnels from each region to the on-premises peer gateway.

D. Deploy Cloud VPN Gateway in each region. Ensure that each region has at least one VPN tunnel to the on-premises peer gateway.

ANSWER: D

Explanation:

Deploy Cloud VPN Gateway in each region. Ensure that each region has at least one VPN tunnel to the on-premises peer gateway. This is correct because Cloud VPN resources are regional: a VPN gateway and its tunnels are associated with a specific Google Cloud region, even though the VPC network itself is global and can contain subnets and Compute Engine workloads in multiple regions. Deploying a gateway in every region that hosts application workloads provides a direct VPN connectivity path for workloads in that region to reach the on-premises network. This architecture also avoids relying on a VPN gateway in one region as the sole regional entry point for workloads located elsewhere.

For production availability, Google recommends using HA VPN, which provides two interfaces on the Cloud VPN gateway and is normally configured with redundant tunnels to appropriately redundant peer VPN infrastructure. The requirement in this scenario is met by establishing regional Cloud VPN connectivity for each application region; the tunnel design can then be extended to HA VPN redundancy as required by the availability objective. See the Cloud VPN overview and Cloud VPN topologies documentation.

QUESTION NO: 22

For this question, refer to the Dress4Win case study.

Dress4Win has asked you to recommend machine types they should deploy their application servers to. How should you proceed?

A. Perform a mapping of the on-premises physical hardware cores and RAM to the nearest machine types in the cloud.

B. Recommend that Dress4Win deploy application servers to machine types that offer the highest RAM to CPU ratio available.

C. Recommend that Dress4Win deploy into production with the smallest instances available, monitor them over time, and scale the machine type up until the desired performance is reached.

D. Identify the number of virtual cores and RAM associated with the application server virtual machines align them to a custom machine type in the cloud, monitor performance, and scale the machine types up until the desired performance is reached.

ANSWER: D

Explanation:

Identify the number of virtual cores and RAM associated with the application server virtual machines, align them to a custom machine type in the cloud, monitor performance, and scale the machine types up until the desired performance is reached. This is the appropriate migration and sizing approach because it begins with the known resource profile of the existing application servers rather than making an arbitrary assumption about memory density or starting production on deliberately undersized instances. Compute Engine custom machine types allow a VM's vCPU and memory configuration to be tailored to workload requirements, which is useful when a predefined machine type does not closely match the existing server profile. Starting with an equivalent, appropriately sized baseline reduces migration risk and helps preserve expected application behavior. After deployment, Cloud Monitoring metrics can validate actual CPU, memory, and other resource consumption under real workload conditions. The machine configuration can then be adjusted based on observed demand and performance targets. This iterative validation is consistent with Google Cloud's rightsizing guidance: use workload metrics to select resources that meet performance requirements while avoiding unnecessary capacity. See the Compute Engine custom machine types documentation and machine type recommendations documentation.

QUESTION NO: 23

The operations manager asks you for a list of recommended practices that she should consider when migrating a J2EE application to the cloud.

Which three practices should you recommend? (Choose three.)

A. Port the application code to run on Google App Engine

B. Integrate Cloud Dataflow into the application to capture real-time metrics

C. Instrument the application with a monitoring tool like Stackdriver Debugger

D. Select an automation framework to reliably provision the cloud infrastructure

E. Deploy a continuous integration tool with automated testing in a staging environment

F. Migrate from MySQL to a managed NoSQL database like Google Cloud Datastore or Bigtable

ANSWER: A D E

Explanation:

Port the application code to run on Google App Engine is a sound migration practice because App Engine provides a managed application platform for Java workloads. The migration should assess and adapt the application for the target

runtime and its managed-service model, rather than simply moving servers without considering operational characteristics. Google documents Java application deployment and runtime configuration for App Engine at App Engine Java runtime.

Select an automation framework to reliably provision the cloud infrastructure is also recommended. Infrastructure as code makes environments repeatable, reviewable, and consistent across development, staging, and production. It reduces manual configuration drift and supports controlled changes during and after the migration. Google recommends using declarative infrastructure management practices and provides tooling guidance at Infrastructure as code with Terraform on Google Cloud.

Deploy a continuous integration tool with automated testing in a staging environment supports safe modernization and release delivery. Automated builds, tests, and a production-like staging environment identify compatibility, integration, and deployment issues before production rollout. This approach enables incremental changes, reliable releases, and rapid feedback while the J2EE application is being adapted to its cloud deployment model.

QUESTION NO: 24

Your company hosts an administrative web application behind an external Application Load Balancer. Only employees in the corporate identity provider should be able to access the application. The security team also wants protection against common web attacks and volumetric HTTP(S) attacks.

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

A. Configure Identity-Aware Proxy (IAP) to protect the application.

B. Attach a Cloud Armor security policy to the external Application Load Balancer.

C. Grant all employees the Compute Admin role on the backend project.

D. Allow TCP port 22 from the internet to each application backend.

E. Replace the external Application Load Balancer with a Cloud VPN tunnel.

ANSWER: A B

Explanation:

Configure Identity-Aware Proxy (IAP) to protect the application. IAP authenticates users and authorizes access using IAM, allowing the company to grant access to the required employee groups without exposing the application to unauthenticated internet users. This provides identity-based access control at the application frontend.

Attach a Cloud Armor security policy to the external Application Load Balancer. Cloud Armor provides DDoS protection and configurable web application firewall controls, including preconfigured rules that can help mitigate common web attacks. Using Cloud Armor with the load balancer protects the application before requests reach the backend. See Identity-Aware Proxy overview and Cloud Armor overview.

QUESTION NO: 25

Your security team needs to retain administrative audit logs for seven years. The logs must be centrally accessible to the security team, protected from modification or deletion during the retention period, and separate from application projects.

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

A. Create a dedicated logging project and use an organization-level log sink to route audit logs to a log bucket in that project.

B. Configure and lock a seven-year retention policy on the destination log bucket.

C. Grant the security team the Owner role on every application project.

D. Export logs manually from each application project to local storage every month.

E. Disable Data Access audit logs to reduce the amount of retained log data.

ANSWER: A B

Explanation:

Create a dedicated logging project and use organization-level log sinks to route the required audit logs to a log bucket in that project. This provides a central security-controlled destination that is separate from the projects generating the logs.

Configure a locked retention policy on the destination log bucket for seven years. Locking the retention policy prevents the retention period from being reduced or removed, helping meet immutable audit-record requirements. The security team should also use least-privilege IAM roles for access to the logging project. See Cloud Logging routing overview and Cloud Logging storage and retention.

QUESTION NO: 26

Your company has an application deployed on Anthos clusters (formerly Anthos GKE) that is running multiple microservices. The cluster has both Anthos Service Mesh and Anthos Config Management configured. End users inform you that the application is responding very slowly. You want to identify the microservice that is causing the delay. What should you do?

A. Use the Service Mesh visualization in the Cloud Console to inspect the telemetry between the microservices.

B. Use Anthos Config Management to create a ClusterSelector selecting the relevant cluster. On the Google Cloud Console page for Google Kubernetes Engine, view the Workloads and filter on the cluster. Inspect the configurations of the filtered workloads.

C. Use Anthos Config Management to create a namespaceSelector selecting the relevant cluster namespace. On the Google Cloud Console page for Google Kubernetes Engine, visit the workloads and filter on the namespace. Inspect the configurations of the filtered workloads.

D. Reinstall istio using the default istio profile in order to collect request latency. Evaluate the telemetry between the microservices in the Cloud Console.

ANSWER: A

Explanation:

Use the Service Mesh visualization in the Cloud Console to inspect the telemetry between the microservices. Anthos Service Mesh provides observability for service-to-service traffic through telemetry collected by the managed service mesh. Its service dashboard and topology visualization show how services communicate and expose metrics such as request volume, error rates, and latency. This lets you follow a slow request through the microservice graph and determine which service or service-to-service edge has elevated latency, then drill into the relevant service metrics for a defined time range.

This approach directly addresses a performance issue in a distributed application without modifying the application or redeploying mesh components. Anthos Config Management is intended to apply and manage configuration consistently across clusters, namespaces, and workloads; it is not the primary mechanism for tracing request latency between running microservices. The mesh visualization uses the traffic telemetry already available from Anthos Service Mesh to identify the location of the delay. Google documents the service mesh observability workflow and the topology view at Anthos Service Mesh observability and View service metrics.

QUESTION NO: 27

Your company has decided to build a backup replica of their on-premises user authentication PostgreSQL database on Google Cloud Platform. The database is 4 TB, and large updates are frequent. Replication requires private address space communication.

Which networking approach should you use?

A. Google Cloud Dedicated Interconnect

B. Google Cloud VPN connected to the data center network

C. A NAT and TLS translation gateway installed on-premises

D. A Google Compute Engine instance with a VPN server installed connected to the data center network

ANSWER: A

Explanation:

Google Cloud Dedicated Interconnect is the appropriate choice because continuous replication of a 4 TB PostgreSQL database with frequent large updates requires high, predictable network throughput and private connectivity between the on-premises environment and the VPC. Dedicated Interconnect provides a direct physical connection from the organization's network to Google's network, avoiding the public internet for traffic to VPC resources. This enables on-premises systems to communicate with private RFC 1918 addresses in the VPC through Cloud Router and BGP route exchange, which meets the private-address-space requirement without placing the database replication flow on public endpoints.

It is designed for sustained, high-volume hybrid-network traffic and offers dedicated connection capacities suitable for workloads whose data-transfer demand is significantly greater than typical VPN use cases. This makes it a strong architectural fit for ongoing database replication, where bandwidth, latency consistency, and operational reliability directly affect replica freshness and recovery objectives. For resilience, production designs should use redundant Dedicated Interconnect connections in separate edge availability domains and configure redundant Cloud Routers and BGP sessions. Google documents Dedicated Interconnect architecture, private IP reachability, and redundancy guidance at Dedicated Interconnect overview and Configure on-premises routers for Dedicated Interconnect.

QUESTION NO: 28

For this question, refer to the EHR Healthcare case study. You are responsible for ensuring that EHR's use of Google Cloud will pass an upcoming privacy compliance audit. What should you do? (Choose two.)

A. Verify EHR's product usage against the list of compliant products on the Google Cloud compliance page.

B. Advise EHR to execute a Business Associate Agreement (BAA) with Google Cloud.

C. Use Firebase Authentication for EHR's user facing applications.

D. Implement Prometheus to detect and prevent security breaches on EHR's web-based applications.

E. Use GKE private clusters for all Kubernetes workloads.

ANSWER: A B

Explanation:

For a healthcare organization handling protected health information, audit readiness requires confirming both the contractual and service-eligibility foundations for HIPAA compliance. "Advise EHR to execute a Business Associate Agreement (BAA) with Google Cloud." is necessary because Google acts as a business associate when it stores, processes, or transmits protected health information on EHR's behalf. Google's HIPAA guidance states that customers must accept a Business Associate Agreement before using Google Cloud services with protected health information.

"Verify EHR's product usage against the list of compliant products on the Google Cloud compliance page." is also essential. A signed agreement does not make every cloud product automatically eligible for regulated healthcare workloads. EHR must ensure that each service used to process protected health information is included in Google Cloud's HIPAA-supported product list and must configure and use those services consistently with its own compliance obligations. This review creates auditable evidence that the architecture uses services covered by Google's HIPAA commitments. Google emphasizes that customers remain responsible for evaluating their particular regulatory obligations and configuring their environments appropriately. See the Google Cloud HIPAA compliance overview and the HIPAA-covered products list.

QUESTION NO: 29

As part of Dress4Win's plans to migrate to the cloud, they want to be able to set up a managed logging and monitoring system so they can handle spikes in their traffic load. They want to ensure that:

* The infrastructure can be notified when it needs to scale up and down to handle the ebb and flow of usage throughout the day * Their administrators are notified automatically when their application reports errors.

* They can filter their aggregated logs down in order to debug one piece of the application across many hosts

Which Google StackDriver features should they use?

A. Logging, Alerts, Insights, Debug

B. Monitoring, Trace, Debug, Logging

C. Monitoring, Logging, Alerts, Error Reporting

D. Monitoring, Logging, Debug, Error Report

ANSWER: C

Explanation:

Monitoring, Logging, Alerts, Error Reporting provides the complete managed observability workflow described. Cloud Monitoring collects infrastructure and application metrics, such as utilization, request rate, and latency. Administrators can define alerting policies with metric thresholds or other conditions and send notifications through configured notification channels. Those alerts can also provide the operational signal used by an autoscaling design to respond to changing demand. Cloud Logging centrally ingests log entries from resources and applications, then enables queries and filters using log fields, resource metadata, labels, and time ranges. This lets operators isolate the behavior of a particular service or application component even when it is distributed across many hosts. Error Reporting automatically groups and analyzes application errors detected in logs, exposing error groups and allowing notifications for newly occurring or recurring failures. Together, these services cover capacity-related monitoring and alerting, centralized log investigation, and automatic visibility into application exceptions. Google's current product names are Cloud Monitoring, Cloud Logging, and Error Reporting, although they were historically grouped under Stackdriver. See the Cloud Monitoring alerting documentation and the Error Reporting documentation.

QUESTION NO: 30

For this question, refer to the Mountkirk Games case study.

Mountkirk Games wants to set up a real-time analytics platform for their new game. The new platform must meet their technical requirements. Which combination of Google technologies will meet all of their requirements?

A. Container Engine, Cloud Pub/Sub, and Cloud SQL

B. Cloud Dataflow, Cloud Storage, Cloud Pub/Sub, and BigQuery

C. Cloud SQL, Cloud Storage, Cloud Pub/Sub, and Cloud Dataflow

D. Cloud Dataproc, Cloud Pub/Sub, Cloud SQL, and Cloud Dataflow

E. Cloud Pub/Sub, Compute Engine, Cloud Storage, and Cloud Dataproc

ANSWER: B

Explanation:

Cloud Dataflow, Cloud Storage, Cloud Pub/Sub, and BigQuery meets the complete set of requirements with fully managed services. Cloud Pub/Sub provides globally scalable, asynchronous ingestion of streaming events directly from game servers. It decouples producers from processing and absorbs variations in game activity without Mountkirk managing messaging infrastructure. Dataflow processes those streams with managed, autoscaling Apache Beam pipelines. Its event-time processing and windowing capabilities are designed to handle late-arriving events, which is essential for telemetry sent over unreliable or slow mobile connections.

Cloud Storage provides durable, highly scalable object storage for files uploaded from players' mobile devices. Those files can be incorporated into the processing pipeline or retained as source data. BigQuery is a serverless analytics data warehouse that supports SQL analysis across very large historical datasets, including the required minimum of 10 TB, while

also accepting streaming data for near-real-time analysis. Together, these services separate ingestion, transformation, file storage, and interactive analytics, while automatically scaling with player activity and avoiding operational responsibility for servers or clusters. See the Google Cloud real-time analytics architecture guidance and the Dataflow streaming with Pub/Sub documentation.

QUESTION NO: 31

Your development team has installed a new Linux kernel module on the batch servers in Google Compute Engine (GCE) virtual machines (VMs) to speed up the nightly batch process. Two days after the installation, 50% of web application deployed in the same

nightly batch run. You want to collect details on the failure to pass back to the development team. Which three actions should you take? Choose 3 answers

A. Use Stackdriver Logging to search for the module log entries.

B. Read the debug GCE Activity log using the API or Cloud Console.

C. Use gcloud or Cloud Console to connect to the serial console and observe the logs.

D. Identify whether a live migration event of the failed server occurred, using in the activity log.

E. Adjust the Google Stackdriver timeline to match the failure time, and observe the batch server metrics.

F. Export a debug VM into an image, and run the image on a local server where kernel log messages will be displayed on the native screen.

ANSWER: A C E

Explanation:

Use Stackdriver Logging to search for the module log entries because centralized logging is the primary place to correlate application, system, and kernel-related messages across the affected batch VMs. Searching by time range, instance, and relevant log content gives the development team durable evidence of errors occurring during the batch execution. Use gcloud or Cloud Console to connect to the serial console and observe the logs because the serial console exposes low-level console output that remains useful when a VM is unhealthy or ordinary network access is unavailable. This can capture boot, kernel, and module diagnostic output that is especially relevant after a kernel-module deployment. Adjust the Google Stackdriver timeline to match the failure time, and observe the batch server metrics to correlate the failure window with operational signals such as CPU utilization, disk I/O, network traffic, and instance behavior. That time-based correlation helps establish the conditions under which the failures began and identifies useful diagnostic context to send to developers. Google Cloud documents serial-console access and its diagnostic role at Troubleshooting using the serial console. Cloud Logging query and analysis capabilities are documented at Logging query language overview.

QUESTION NO: 32

For this question, refer to the Cymbal Retail case study Cymbal plans to migrate their existing on-premises systems to Google Cloud and implement AI-powered virtual agents to handle customer interactions You need to provision the compute resources that can scale for the AI-powered virtual agents What should you do?

A. Use Cloud SQL to store the customer data and product catalog.

B. Configure Cloud Build to call AI Applications (formerly Vertex AI Agent Builder).

C. Deploy a Google Kubernetes Engine (GKE) cluster with autoscaling enabled

D. Create a single, large Compute Engine VM instance with a high CPU allocation.

ANSWER: C

Explanation:

Deploy a Google Kubernetes Engine (GKE) cluster with autoscaling enabled is the appropriate choice because the virtual-agent application needs a compute platform that can adjust capacity as customer demand varies. GKE runs containerized workloads as replicated Pods, allowing the agent service to be deployed as a resilient, horizontally scalable application rather than being limited to the capacity of one server. This is particularly useful for retail workloads, where customer interactions can increase sharply during promotions, seasonal events, or other peak periods.

GKE autoscaling addresses both application and infrastructure capacity. The Horizontal Pod Autoscaler can increase or decrease the number of Pods based on observed resource utilization or custom metrics, so more agent-service instances can process requests when load rises. Cluster autoscaling can then add or remove nodes when the Pod capacity required by the workload changes. Together, these mechanisms provide elastic compute capacity while avoiding the need to maintain peak-sized resources continuously. This aligns with a migration architecture that needs availability, operational flexibility, and cost-conscious scaling for containerized services. Google documents GKE autoscaling options at GKE autoscaling and explains workload-level scaling through the Horizontal Pod Autoscaler.

QUESTION NO: 33

For this question, refer to the EHR Healthcare case study. You need to define the technical architecture for securely deploying workloads to Google Cloud. You also need to ensure that only verified containers are deployed using Google Cloud services. What should you do? (Choose two.)

A. Enable Binary Authorization on GKE, and sign containers as part of a CI/CD pipeline.

B. Configure Jenkins to utilize Kritis to cryptographically sign a container as part of a CI/CD pipeline.

C. Configure Container Registry to only allow trusted service accounts to create and deploy containers from the registry.

D. Configure Container Registry to use vulnerability scanning to confirm that there are no vulnerabilities before deploying the workload.

ANSWER: A B

Explanation:

Enabling Binary Authorization on GKE and signing containers in the CI/CD pipeline establishes an enforceable deployment gate for container workloads. Binary Authorization evaluates configured policy before GKE admits an image, allowing the organization to require a valid attestation for each deployed container. This provides a cryptographic chain of trust: an image is built, reviewed or tested by the delivery process, signed, and then permitted to run only when its required attestation is present and valid. This control is particularly appropriate for healthcare environments because it prevents unverified images from bypassing the approved software-delivery process.

Configuring Jenkins to use Kritis to cryptographically sign a container is the complementary attestation-generation step. Kritis is designed to create and validate image attestations used with Binary Authorization policy enforcement. Integrating signing into the CI/CD workflow ensures that verified build artifacts receive attestations consistently and automatically, rather than relying on an administrator to approve deployments manually. Together, signing through the pipeline and admission enforcement in GKE ensure that only containers satisfying the organization's verification policy can be deployed. Google documents Binary Authorization policy enforcement and attestation requirements at Binary Authorization overview and describes the Kritis attestation framework at Kritis on GitHub.

QUESTION NO: 34

Refer to the **Altostrat Media** case study for the following solution.

Altostrat is concerned about sophisticated, multi-vector Distributed Denial of Service (DDoS) attacks targeting various layers of their infrastructure. DDoS attacks could potentially disrupt video streaming and cause financial losses. You need to mitigate this risk. What should you do?

A. Set up VPC Service Controls to restrict access to sensitive resources and prevent data exfiltration.

B. Configure Cloud Next Generation Firewall (NGFW) with custom rules to filter malicious traffic at the network level.

C. Deploy Google Cloud Armor with pre-configured and custom rules for L3/L4 and L7 protection.

D. Activate Security Command Center to monitor security posture and detect potential threats.

ANSWER: C

Explanation:

Deploy Google Cloud Armor with pre-configured and custom rules for L3/L4 and L7 protection. This approach directly addresses a multi-vector DDoS threat by combining protections designed for both network-layer volumetric attacks and application-layer attacks against the video-streaming service. Cloud Armor's network DDoS protection is designed to detect and mitigate packet-based attacks at the edge before they can consume backend capacity or disrupt services. For HTTP(S) workloads behind supported Google Cloud load balancers, Cloud Armor security policies provide application-layer controls, including custom allow or deny rules, rate limiting, adaptive protection, and preconfigured Web Application Firewall rules. These controls can be tailored to known traffic patterns, abusive source characteristics, and threats directed at application endpoints. Applying the protections at Google's edge is particularly valuable for a streaming platform because malicious traffic can be absorbed and filtered before it reaches the application infrastructure, preserving capacity for legitimate viewers. Google documents its layered DDoS capabilities in the Cloud Armor DDoS protection overview and describes HTTP(S) security policies and rule enforcement in the Cloud Armor security policy overview.

QUESTION NO: 35

For this question, refer to the Mountkirk Games case study. Mountkirk Games wants to design their solution for the future in order to take advantage of cloud and technology improvements as they become available. Which two steps should they take? (Choose two.)

A. Store as much analytics and game activity data as financially feasible today so it can be used to train machine learning models to predict user behavior in the future.

B. Begin packaging their game backend artifacts in container images and running them on Kubernetes Engine to improve the availability to scale up or down based on game activity.

C. Set up a CI/CD pipeline using Jenkins and Spinnaker to automate canary deployments and improve development velocity.

D. Adopt a schema versioning tool to reduce downtime when adding new game features that require storing additional player data in the database.

E. Implement a weekly rolling maintenance process for the Linux virtual machines so they can apply critical kernel patches and package updates and reduce the risk of 0-day vulnerabilities.

ANSWER: B C

Explanation:

Packaging game-backend artifacts as container images and running them on Google Kubernetes Engine is a future-oriented architecture choice because containers provide a consistent, portable unit of deployment. GKE supplies managed Kubernetes capabilities while allowing Mountkirk Games to deploy updated services independently, adjust capacity as player demand changes, and benefit from ongoing improvements to the managed Kubernetes platform. This approach reduces coupling between application deployment and the underlying virtual-machine configuration, which makes later technology adoption less disruptive. Google documents GKE's managed, scalable Kubernetes environment at Google Kubernetes Engine overview.

Establishing a CI/CD pipeline with Jenkins and Spinnaker also prepares the organization to adopt future technologies safely and quickly. Automated build, test, and delivery workflows create a repeatable path for releasing backend changes. Canary deployments are especially valuable for a live game platform because they progressively expose a release to a limited portion of traffic, provide an opportunity to observe production behavior, and support controlled rollout of new functionality. A mature continuous-delivery process therefore improves development velocity while making future platform, application, and service changes easier to validate and deploy. Google's guidance on deployment strategies, including canary releases, is available at Application deployment and testing strategies.

QUESTION NO: 36

Your company wants to migrate their 10-TB on-premises database export into Cloud Storage. You want to minimize the time it takes to complete this activity, the overall cost, and database load. The bandwidth between the on-premises environment and Google Cloud is 1 Gbps. You want to follow Google-recommended practices. What should you do?

A. Develop a Dataflow job to read data directly from the database and write it into Cloud Storage.

B. Use the Data Transfer appliance to perform an offline migration.

C. Use a commercial partner ETL solution to extract the data from the on-premises database and upload it into Cloud Storage.

D. Compress the data and upload it with gsutil -m to enable parallel copy operations.

ANSWER: D

Explanation:

Compress the data and upload it with gsutil -m to enable parallel copy operations is the appropriate approach because the source is already a database export. The migration can therefore operate on export files rather than repeatedly reading from the live database, keeping additional database load low. Compression reduces the number of bytes that must traverse the 1 Gbps connection, which can materially reduce transfer duration when the exported data is compressible. A 10 TB uncompressed transfer at 1 Gbps has a theoretical minimum of roughly 22 hours before protocol overhead, retries, and operational delays; reducing the transferred size helps improve this practical completion time.

Using `gsutil -m` runs independent copy operations in parallel, allowing multiple exported files to use available network capacity effectively. This is a direct, managed-service-compatible upload approach and does not add the procurement, shipping, or third-party platform costs associated with alternative transfer methods. Cloud Storage uploads are supported through gsutil, and Google documents the `-m` flag for parallel operations. For large local data transfers, Google also recommends planning uploads around file organization and parallelism so that the available connection is efficiently used. See the gsutil cp command documentation and Cloud Storage upload documentation.

QUESTION NO: 37

A lead engineer wrote a custom tool that deploys virtual machines in the legacy data center. He wants to migrate the custom tool to the new cloud environment. You want to advocate for the adoption of Google Cloud Deployment Manager.

What are two business risks of migrating to Cloud Deployment Manager? (Choose two.)

A. Cloud Deployment Manager uses Python

B. Cloud Deployment Manager APIs could be deprecated in the future

C. Cloud Deployment Manager is unfamiliar to the company's engineers

D. Cloud Deployment Manager requires a Google APIs service account to run

E. Cloud Deployment Manager can be used to permanently delete cloud resources

F. Cloud Deployment Manager only supports automation of Google Cloud resources

ANSWER: B C

Explanation:

Cloud Deployment Manager APIs could be deprecated in the future is a material business risk because infrastructure deployment tooling is foundational to operations. A product lifecycle change can require the organization to fund, plan, test, and execute a migration to a replacement infrastructure-as-code platform. This risk is especially concrete for Deployment Manager: Google has announced that Deployment Manager support ends on March 31, 2026. Adopting it for a new migration would therefore create a short-lived dependency and potentially duplicate migration effort. Google recommends moving deployments to Infrastructure Manager or another supported IaC solution. See the Deployment Manager deprecation notice.

Cloud Deployment Manager is unfamiliar to the company's engineers is also a business risk. Adoption requires engineers to learn its configuration syntax, templates, deployment lifecycle, permissions model, troubleshooting procedures, and operational controls. That learning and enablement work can slow delivery, increase early implementation errors, and create a key-person or supportability risk if only a small group becomes proficient. The organization should account for training, documentation, code review standards, and operational handover when evaluating an infrastructure automation platform. Google's Deployment Manager documentation describes the product's templates and deployment concepts that engineering teams would need to operate.

QUESTION NO: 38

Your operations team needs to detect failures in a public API before customers report them. The API runs behind an external Application Load Balancer, and the team needs alerts when availability falls below the service objective and when request latency exceeds the target.

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

A. Create an availability service level indicator and service level objective based on successful API requests.

B. Create a latency service level indicator and service level objective based on requests completed within the latency target.

C. Use only VM CPU utilization as the service level indicator.

D. Alert only when the load balancer has no configured backend instances.

E. Configure a monthly manual review of application logs without alerting policies.

ANSWER: A B

Explanation:

Create an availability service level indicator based on successful API requests and define an availability service level objective. An availability SLI measures the proportion of requests that are successfully served from the user's perspective. Cloud Monitoring can calculate error-budget consumption and alert when the service is at risk of missing its availability objective.

Create a latency service level indicator based on the proportion of requests completed within the latency target, and define a latency service level objective. This measures whether users receive timely responses rather than only whether the service responds at all. Alerting on SLO burn rate or objective risk gives the operations team an actionable signal before a full outage occurs. See Service monitoring and SLOs and Cloud Monitoring alerting.

QUESTION NO: 39

For this question, refer to the JencoMart case study.

The migration of JencoMart's application to Google Cloud Platform (GCP) is progressing too slowly. The infrastructure is shown in the diagram. You want to maximize throughput. What are three potential bottlenecks? (Choose 3 answers.)

A. A single VPN tunnel, which limits throughput

B. A tier of Google Cloud Storage that is not suited for this task

C. A copy command that is not suited to operate over long distances

D. Fewer virtual machines (VMs) in GCP than on-premises machines

E. A separate storage layer outside the VMs, which is not suited for this task

F. Complicated internet connectivity between the on-premises infrastructure and GCP

ANSWER: A C D

Explanation:

A single VPN tunnel, which limits throughput, is a likely bottleneck because Cloud VPN has per-tunnel bandwidth limits. A migration flow that uses only one tunnel has a fixed upper bound on the amount of traffic it can move, regardless of the capacity available in the source or destination environment. Cloud VPN supports multiple tunnels and Equal-Cost Multi-Path routing to increase aggregate available bandwidth when the design and routing configuration support it.

A copy command that is not suited to operate over long distances can also constrain the transfer. Long round-trip times reduce the effectiveness of unoptimized, single-stream copy operations. Migration tooling should support efficient parallel transfer, retries, and resumability so that the transfer can better use available WAN bandwidth. Google Cloud Storage provides parallelized transfer capabilities for large-scale object movement.

Fewer virtual machines (VMs) in GCP than on-premises machines can limit throughput when the migrated application relies on horizontally scaled workers, application servers, or processing nodes. The target environment needs enough compute capacity and parallelism to receive, process, and serve the migrated workload at the required rate. See Cloud VPN bandwidth considerations and Cloud Storage parallel composite uploads.

QUESTION NO: 40

Your company is developing a web-based application. You need to make sure that production deployments are linked to source code commits and are fully auditable. What should you do?

A. Make sure a developer is tagging the code commit with the date and time of commit.

B. Make sure a developer is adding a comment to the commit that links to the deployment.

C. Make the container tag match the source code commit hash.

D. Make sure the developer is tagging the commits with latest.

ANSWER: C

Explanation:

Make the container tag match the source code commit hash. A commit SHA is a unique identifier for a specific version of source code, so using it as the container image tag establishes a direct, machine-verifiable relationship between the artifact deployed to production and the exact commit that produced it. Deployment records can then identify the image tag (and, ideally, its immutable image digest), while the source repository identifies the associated code, author, review history, and build context. This supports reliable traceability during incident response, release reviews, and compliance audits without relying on manual developer actions.

Cloud Build provides the `COMMIT_SHA` built-in substitution for builds triggered from a repository, enabling build configurations to tag images consistently with the commit SHA. Google recommends recording immutable artifact references such as image digests for deployment provenance; pairing that immutable reference with a commit-SHA tag makes the source-to-production trail clear and repeatable. See Cloud Build default substitutions and Cloud Deploy image deployment documentation.