Google Cloud Professional Cloud Architect Practice Exam Questions (2024-2025)

The October 2025 update (Version 6.1) significantly transformed this certification with **new AI/ML sections**, mandatory **Well-Architected Framework** knowledge, and **three new case studies** replacing legacy scenarios. gcpstudyhub Shinglyu The exam now comprises **50-60 questions over 2 hours**, with **20-30% based on case studies** Google Cloud that all incorporate AI integration requirements. GCP Study Hub

Critical October 2025 updates you must know

New exam sections 2.4 and 2.5 specifically cover Vertex AI for end-to-end ML workflows and pre-built AI solutions. GCP Study Hub The **Google Cloud Well-Architected Framework** is now mandatory knowledge woven throughout all objectives. Google Cloud gcpstudyhub New security topics include **Model Armor** for AI security and **Sensitive Data Protection**. Shinglyu

The case studies have been completely refreshed: **Altostrat Media** (media company with Gen AI for content analysis), **Cymbal Retail** (online retailer using Gen AI for catalog management), **EHR Healthcare** (retained from previous exam), and **KnightMotives Automotive** (new). TerramEarth, Mountkirk Games, and Helicopter Racing League have been removed. GCP Study Hub +2

Section 1: Designing and planning cloud solution architecture (~25%)

Question 1: Data ingestion architecture

A vehicle telematics company receives daily data using network interconnects with private on-premises data centers. A subset of data is transmitted and processed in real time; the rest arrives daily when vehicles return to base. You need a complete solution for ingestion and management, fully storing data and aggregating for BigQuery analytics.

Which actions are best? (Choose 2)

A. Real-time data is streamed directly to BigQuery, and daily jobs create aggregate processing
B. Real-time data is sent via Pub/Sub and processed by Dataflow that stores data in Cloud Storage and computes aggregates for BigQuery
C. Daily sensor data is uploaded to Cloud Storage with parallel composite uploads; a Cloud Storage trigger activates Dataflow
D. Daily sensor data is loaded with BigQuery Data Transfer Service and processed on demand

Correct Answer: B, C

Pub/Sub provides reliable, many-to-many asynchronous messaging with at-least-once delivery through its loosely coupled publish/subscribe mechanism. Whizlabs Dataflow manages both real-time streaming and batch

processing using the same Apache Beam procedures. **Parallel composite uploads** are specifically recommended for large files in the **200-500 MB** range, and Cloud Storage triggers can automatically activate Dataflow procedures. whizlabs

Option A stores data only in BigQuery without real-time processing capabilities. Option D fails because BigQuery Data Transfer Service handles cloud-to-cloud transfers, not on-premises data, and cannot handle the required decompression and processing. whizlabs

---

Question 2: API versioning strategy

Your company is making a major API revision to improve developer experience. They must keep the old version available and deployable while allowing new customers to test the new API. They want to maintain the same SSL certificate and DNS records to serve both versions. ExamTopics

What should they do?

A. Configure a new load balancer for the new version
B. Reconfigure old clients to use a new endpoint for the new API
C. Have the old API forward traffic to the new API based on the path
D. Use separate backend pools for each API path behind the load balancer

Correct Answer: D

Using **separate backend pools for each API path** behind the HTTP(S) load balancer enables both API versions to coexist using identical SSL certificates and DNS records. URL path-based routing directs traffic to different backend services based on request paths (e.g., /v1/* to the legacy backend, /v2/* to the new backend), enabling seamless versioning without DNS changes or certificate reissuance.

Question 3: Multi-petabyte analytics storage

Your company plans to migrate a multi-petabyte dataset to the cloud. The data must be available 24/7. Your business analysts have experience only with SQL interfaces.

How should you store the data to optimize for ease of analysis?

A. Load data into Google BigQuery
B. Insert data into Google Cloud SQL
C. Put flat files into Google Cloud Storage
D. Stream data into Google Cloud Datastore

Correct Answer: A

BigQuery is purpose-built for multi-petabyte scale analytics with a familiar SQL interface and 24/7 availability. It uses columnar storage and automatic sharding for query performance at scale. Cloud SQL has storage limitations making petabyte-scale impractical. Cloud Storage provides object storage without native SQL query capabilities. Datastore is a NoSQL document database without SQL support.

Question 4: High availability for REST APIs

Which Compute Engine architecture provides autoscaling, global low latency, and multi-zone high availability for a REST API experiencing traffic spikes up to 120,000 requests per second?

A. Zonal managed instance group per region with external HTTPS load balancing
B. Regional managed instance groups behind an external TCP proxy
C. Regional autoscaled MIGs per region with a global external HTTPS load balancer

Correct Answer: C

This architecture delivers all three requirements: **Autoscaling** that reacts to sudden traffic spikes because managed instance groups scale based on CPU, HTTP load balancing utilization, or custom metrics. **Global low latency** because the global external HTTPS load balancer uses anycast to route users to the nearest healthy backend. **High availability** because regional MIGs distribute instances across multiple zones within each region, surviving single-zone failures.

Question 5: Zone failure resilience

To ensure your application handles the load even if an entire zone fails, what should you do? (Select all that apply)

A. Don't select the "Multizone" option when creating your managed instance group
B. Spread your managed instance group over two zones and overprovision by 100%
C. Create a regional unmanaged instance group and spread instances across multiple zones
D. Overprovision your regional managed instance group by at least 50% for three zones

Correct Answer: B, D

With **100% overprovisioning across two zones**, each zone contains 100% of desired capacity. If one fails, you retain 100% capacity. With **50% overprovisioning across three zones**, each zone has 50% of desired capacity; losing one zone leaves you with 100% capacity in the remaining two. (whizlabs)

Option C fails because unmanaged instance groups don't auto-scale—they cannot automatically handle the additional load when a zone fails. Google's documentation recommends at least 50% overprovisioning for regional MIGs spanning three or more zones. (whizlabs)

Section 2: Managing and provisioning cloud solution infrastructure (~17.5%)

Question 6: Secure VM management without public IPs

You manage many GCP Compute Engine instances using SSH and RDP. Management requires that VMs cannot have public IP addresses. How can you securely manage access?

A. Bastion Hosts
B. NAT Instances
C. IAP's TCP forwarding
D. Security Command Center

Correct Answer: C

Identity-Aware Proxy (IAP) TCP forwarding enables SSH and RDP connections to VMs without public IP addresses. IAP wraps traffic in HTTPS and validates user access through IAM policies. A proxy server inside GCP translates communication, providing secure access without exposing VMs publicly. whizlabs Whizlabs

Bastion hosts and NAT instances both require public IP addresses, violating the security requirement. Security Command Center monitors vulnerabilities and threats but doesn't provide access management capabilities. Whizlabs

---

Question 7: Instance template updates for MIGs

In a Compute Engine managed instance group, how can you apply a new instance template only to future VMs while keeping existing instances unchanged? (Choose 2)

A. Set PROACTIVE rollout with RECREATE
B. Use OPPORTUNISTIC updates and let autoscaling add instances
C. Create a new managed instance group and shift traffic with a load balancer
D. Use OPPORTUNISTIC updates and manually resize the group

Correct Answer: B, D

With **OPPORTUNISTIC updates**, the MIG accepts the new template but doesn't proactively restart or recreate existing VMs. New instances created through autoscaling or manual resize operations use the new template while existing instances remain unchanged.

PROACTIVE rollout with RECREATE would immediately replace existing instances with new ones using the updated template, violating the requirement to keep existing instances unchanged.

Question 8: Fast-scaling MIG with pre-installed packages

A company plans to automate deployment of a Compute Engine MIG for a latency-sensitive service. Each VM requires numerous OS packages. During peak events, the group scales from 8 to 160 instances and must make new VMs ready quickly.

What should you implement?

A. Use Google Cloud OS Config guest policies to install packages after instances start
B. Build a custom Compute Engine image with all required OS packages and use Deployment Manager to create the MIG
C. Provision the MIG with Terraform and use a startup script to download and install packages
D. Use Puppet to configure instances after MIG creation

Correct Answer: B

Pre-baking all required packages into a custom image eliminates time-consuming installation steps during instance boot. New VMs start from an image that already contains everything needed, minimizing initialization time and supporting fast scale-out for latency-sensitive workloads.

Options A, C, and D all incur package installation at boot time, adding significant delay to instance readiness during sudden scaling events when rapid response is critical.

Question 9: Storage class selection

What is the command for creating a Cloud Storage bucket with once-per-month access frequency named 'archive_bucket'?

A. gsutil rm -coldline gs://archive_bucket
B. gsutil mb -c coldline gs://archive_bucket
C. gsutil mb -c nearline gs://archive_bucket
D. gsutil mb gs://archive_bucket

Correct Answer: C

The `mb` command creates buckets. **Nearline storage** is designed for data accessed approximately once per month, with a **30-day minimum storage duration**. Coldline storage is for data accessed once per 90 days and would incur retrieval charges for monthly access patterns. whizlabs The command `gsutil mb -c nearline gs://archive_bucket` creates the appropriate bucket.

Question 10: App Engine to on-premises MySQL connectivity

Which configuration lets App Engine Standard reach an on-premises MySQL database over private IP

through an existing Cloud VPN?

A. Private Service Connect
B. Serverless VPC Access connector
C. Private Google access
D. Private services access

Correct Answer: B

A **Serverless VPC Access connector** allows App Engine Standard to send traffic into your VPC network. That traffic can then traverse your existing Cloud VPN tunnel to reach on-premises MySQL over private IP addresses.

Private Google Access enables access to Google APIs without external IPs. Private services access is for connecting to Google-managed services like Cloud SQL. Private Service Connect provides private endpoints for Google services, not on-premises connectivity.

Section 3: Designing for security and compliance (~17.5%)

Question 11: Organization-wide external IP restrictions

Following a cyber incident, the CISO requires a control that prevents production Compute Engine VMs from obtaining external IP addresses unless explicitly approved. You need a straightforward, centrally enforced solution.

What should you do?

A. Use VPC Service Controls to place production projects in a service perimeter
B. Replace default internet gateway routes with Cloud NAT
C. Apply the Organization Policy constraint constraints/compute.vmExternalIpAccess at the organization or folder level
D. Build two custom VPC networks with different routing configurations

Correct Answer: C

The **Organization Policy constraint** `constraints/compute.vmExternalIpAccess` provides a centrally enforced guardrail that denies external IP assignment by default. It supports narrow allowlists for specific approved VM instances, is auditable and reversible, and can be applied at organization or folder level for comprehensive coverage.

VPC Service Controls protect access to Google-managed APIs, not external IP assignment. Cloud NAT enables outbound connectivity but doesn't prevent users from assigning external IPs. Custom VPC networks add operational complexity without preventing external IP assignment.

Question 12: Secure on-premises API exposure

A healthcare company has legacy API integrations with on-premises systems. They want to use these APIs from Google Cloud applications while keeping them private—exposed securely without direct internet access.

Which technique fulfills these requirements?

A. Gated Egress and VPC Service Controls
B. Cloud Endpoints
C. Cloud VPN
D. Cloud Composer

Correct Answer: A

Gated egress topology makes APIs in on-premises environments available only to processes inside Google Cloud without direct public internet access. Applications communicate with APIs using private IP addresses through an internal Application Load Balancer. whizlabs

VPC Service Controls add security by isolating services and data, monitoring against data exfiltration, and restricting access to authorized IPs and client contexts. Cloud Endpoints doesn't support on-premises endpoints. Cloud VPN connects networks but doesn't provide API-level protection or access control. whizlabs

---

Question 13: Web vulnerability scanning

During testing, developer code allows user input to modify the application and execute commands. You suspect other vulnerabilities exist.

Which service helps identify them?

A. Cloud Armor
B. Web Security Scanner
C. Security Command Center
D. Shielded GKE nodes

Correct Answer: B

Web Security Scanner performs managed and custom web vulnerability scanning, identifying issues like cross-site scripting (XSS), SQL injection, and the code injection vulnerability described. It performs scans for OWASP Top 10, CIS GCP Foundation, and PCI-DSS findings. whizlabs

Cloud Armor is a network security service providing WAF rules and DDoS defense, not application vulnerability scanning. Whizlabs Security Command Center is a broader security suite that contains Web

Security Scanner among other services. Shielded GKE nodes provide secure boot and integrity verification for VMs, not application scanning. Whizlabs

---

Question 14: DDoS protection for HTTPS load balancer

A digital media company experienced DDoS attacks against their HTTPS load balancer. Which service provides defense against DDoS attacks?

A. Cloud Armor
B. Cloud Identity-Aware Proxy
C. GCP Firewalls
D. IAM policies

Correct Answer: A

Cloud Armor delivers defense at scale against infrastructure and application DDoS attacks using Google's global infrastructure. It integrates with HTTP(S) Load Balancing to provide edge-based protection, blocking attacks close to their source with custom security policies and IP allowlisting/denylisting. whizlabs

VPC firewall rules don't apply to HTTP(S) Load Balancer traffic. Cloud IAP provides application-level access control, not DDoS mitigation. IAM policies control resource access permissions, not network attack prevention. whizlabs

---

Question 15: Firewall rule segmentation

When creating firewall rules, what forms of segmentation can narrow which resources the rule is applied to? (Choose all that apply)

A. Network range in source filters
B. Zone
C. Region
D. Network tags

Correct Answer: A, D

VPC firewall rules use **network tags** and **network ranges (CIDR blocks)** for traffic filtering. You can target instances by tags and specify source/destination IP ranges in rules. whizlabs

Zones and regions are not valid selectors for firewall rules. Firewall rules are global resources that apply based on network, tags, service accounts, and IP ranges—not geographic location.

Question 16: Environment isolation for production security

Development and QA teams collaborate and need access to each other's environments, but they can also access staging and production, which raises security risks. Staging must copy production data every 36 hours.

How should you restructure to keep production isolated while allowing required data transfer?

A. Place dev/test in a single VPC, staging/production in a different VPC
B. Create one project for dev/test together, separate projects for staging and production
C. Keep all in one project with VPC Service Controls around production data
D. Deploy dev/test to one subnet, staging/production to another subnet

Correct Answer: B

Projects are the primary isolation boundary for IAM, policies, quotas, and network scope. whizlabs Placing staging and production in separate projects prevents broad access overlap. Grant a dedicated service account in staging narrowly scoped read access to specific production datasets, scheduling transfers every 36 hours with Cloud Scheduler triggering Storage Transfer Service or Dataflow.

VPCs and subnets provide network isolation but not administrative or IAM isolation. Grouping staging with production would violate the isolation requirement.

Section 4: Analyzing and optimizing technical and business processes (~15%)

Question 17: J2EE migration best practices

The operations manager asks for recommended practices when migrating a J2EE application to the cloud.

Which three practices should you recommend? (Choose 3)

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 Cloud 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 Cloud Datastore

Correct Answer: C, D, E

Monitoring instrumentation (C) with Cloud Debugger and Cloud Operations provides essential visibility into application performance and behavior in the cloud environment. **Infrastructure automation (D)** ensures

reliable, repeatable deployments and reduces configuration drift. **CI/CD with automated testing (E)** reduces deployment risks and accelerates release cycles.

Porting to App Engine (A) isn't necessary for migration. Dataflow (B) is for data processing pipelines, not application metrics. Database migration to NoSQL (F) introduces unnecessary complexity unless specifically required.

Question 18: Cross-service latency diagnosis

A fintech company sees sporadic slowdowns when confirming card charges across multiple microservices. Engineers believe interservice latency is the cause.

How should they identify which services add extra time?

A. Enable Cloud Logging for every backend service
B. Turn on Cloud Profiler to analyze CPU and memory usage
C. Instrument with Cloud Trace and view distributed traces
D. Place an external HTTP(S) Load Balancer in front of services

Correct Answer: C

Cloud Trace provides distributed tracing that stitches together spans across all microservice hops. Each hop is timed and visualized on a timeline, showing exactly where latency accumulates. You can drill into spans for individual RPCs and HTTP calls to identify the slow backend service.

Cloud Logging doesn't automatically correlate entire request paths or provide precise per-hop timing. Cloud Profiler focuses on code-level CPU and memory analysis, not network latency between services. Load balancers don't provide visibility into downstream service chain latency.

Question 19: Budget alerting

Your company has reserved a monthly budget. You want to be informed automatically when approaching the limit.

What should you do?

A. Link a credit card with a monthly limit equal to your budget
B. Create a budget alert for desired percentages such as 50%, 90%, and 100% of your total monthly budget
C. In App Engine Settings, set a daily budget at the rate of 1/30 of monthly budget
D. In GCP Console, configure billing export to BigQuery with a saved view

Correct Answer: B

Budget alerts notify stakeholders proactively when spending reaches configured thresholds. Setting alerts at 50%, 90%, and 100% provides early warning and escalation as spending approaches limits. whizlabs

Credit card limits (A) don't provide proactive notifications. App Engine daily budgets (C) only affect App Engine resources and cause requests to fail when exceeded. Billing export to BigQuery (D) provides historical data analysis but not automatic alerting when approaching limits. whizlabs

---

Question 20: Idle VM detection

You've been billed for a large number of Compute Engine instances, which you consider excessive. How can you identify systems accidentally left active?

A. Use the Recommender CLI command
B. Use Cloud Billing Reports
C. Use Idle Systems Report in GCP Console
D. Use Security Command Center Reports

Correct Answer: A

The **Recommender CLI** command `gcloud recommender recommendations list --recommender=google.compute.instance.IdleResourceRecommender` identifies idle VMs based on Cloud Monitoring metrics from the previous 14 days. whizlabs It surfaces VMs with consistently low CPU utilization that may be candidates for downsizing or deletion.

Cloud Billing Reports show spending but not activity levels. There is no "Idle Systems Report" in the GCP Console. Security Command Center focuses on security threats, not operational efficiency. whizlabs

---

Section 5: Managing implementation (~12.5%)

Question 21: Low-risk App Engine deployment

You need to deploy a risky update to an application in Google App Engine. The update can only be tested in a live environment.

What's the best way to minimize risk?

A. Deploy a new version and use traffic splitting to direct only a small number of users to the new version
B. Deploy temporarily and be prepared to pull it back
C. Warn users about potential issues and provide contact methods
D. Create a new project with the new version, then redirect users

Correct Answer: A

Deploying a new version without assigning it as the default version creates no downtime. **Traffic splitting** enables gradual rollout by directing a small percentage of traffic to the new version. If issues arise, traffic can be quickly redirected back to the stable version without application downtime. whizlabs

Option B exposes all users to risk. Option C isn't a technical mitigation strategy. Option D requires unnecessary data synchronization and external traffic routing configuration. Whizlabs

---

Question 22: Cloud Run canary deployment

You manage a Cloud Run service and need to deploy a new version with about 20% of real user traffic to validate performance.

How should you test?

A. Create a parallel Cloud Run service and place an external HTTPS load balancer in front
B. Configure Traffic Director with a new service entry and small weighting
C. Deploy new version as a revision and use traffic splitting to route a small percentage
D. Set up Cloud Build trigger with TRAFFIC_PERCENT substitution variable

Correct Answer: C

Cloud Run supports **revision-based traffic splitting** natively. Deploy the new version as a revision without making it default, then configure percentage-based traffic splitting to send 20% to the new revision. Scribd You can adjust percentages up or down and quickly roll back by changing traffic configuration.

No external load balancer or Traffic Director configuration is needed—Cloud Run handles revision traffic splitting internally. Cloud Build manages build automation, not live traffic routing.

Question 23: Workflow integration for legacy and modern systems

A company deployed 5G devices in vehicles for real-time data transmission, but older vehicles still use legacy technology where data is downloaded via maintenance port. You need to integrate this old procedure with the new one simply.

A. Cloud Composer
B. Cloud Interconnect
C. App Engine
D. Cloud Build

Correct Answer: A

Cloud Composer is a fully managed workflow orchestration service based on Apache Airflow. It can author, schedule, and monitor pipelines spanning cloud and on-premises data centers, making it ideal for integrating

legacy batch data uploads with modern streaming ingestion. whizlabs

Cloud Interconnect provides network connectivity, not workflow orchestration. App Engine requires custom application development. Cloud Build automates code builds and deployments, not data workflow orchestration.

Section 6: Ensuring solution and operations reliability (~12.5%)

Question 24: IoT streaming pipeline architecture

You need to take streaming data from thousands of IoT devices, ingest it, run it through a processing pipeline, and store it for SQL analysis.

What services and order should you use?

A. Cloud Dataflow, Cloud Pub/Sub, BigQuery
B. Cloud Pub/Sub, Cloud Dataflow, Cloud Dataproc
C. Cloud Pub/Sub, Cloud Dataflow, BigQuery
D. App Engine, Cloud Dataflow, BigQuery

Correct Answer: C

The standard GCP streaming architecture follows this pattern: **Cloud Pub/Sub** reliably ingests event streams with at-least-once delivery. **Cloud Dataflow** processes and transforms data with exactly-once semantics. **BigQuery** provides data warehousing with SQL analytics capabilities. whizlabs

Option A has incorrect order (Pub/Sub must come before Dataflow). Option B uses Dataproc (Hadoop/Spark) instead of BigQuery for analytics. Option D uses App Engine, which isn't designed for IoT data ingestion at scale.

Question 25: Mission-critical database data loss minimization

Your team is deploying a Cloud SQL for MySQL instance storing mission-critical payment records. Leadership wants to minimize data loss if a regional outage occurs.

Which features should you enable? (Choose 2)

A. Semisynchronous replication
B. Automated backups
C. Sharding
D. Binary logging
E. Read replicas

Correct Answer: B, D

Automated backups create consistent base backups at regular intervals. **Binary logging** records every change after the last backup, enabling **point-in-time recovery (PITR)** by restoring the last backup and replaying binary logs to a precise timestamp. This combination minimizes potential data loss to the smallest practical window.

Read replicas use asynchronous replication and can lag behind, potentially losing recent transactions. Sharding improves scalability, not recovery capabilities. Semisynchronous replication is not a managed feature for Cloud SQL MySQL.

Question 26: Long-term log retention for compliance

How should you centralize raw logs from all Google Cloud projects and retain them for 10 years in a simple, cost-effective way for auditor access?

A. Export logs from all projects to BigQuery
B. Use aggregated sinks to export all logs to Cloud Storage
C. Stream all logs to Pub/Sub
D. Centralize logs in a Cloud Logging bucket with custom retention

Correct Answer: B

An **aggregated sink** at the organization or folder level captures logs from all descendant projects and routes them to a single Cloud Storage bucket. Cloud Storage supports 10-year retention policies with optional retention lock for immutability, Archive storage class for minimal cost, and straightforward IAM permissions for auditor access.

BigQuery is more expensive for decade-long archival of infrequently accessed data. Pub/Sub is a messaging service, not long-term storage. Cloud Logging buckets cost more than Cloud Storage over long durations.

Question 27: Shared VPC for multinational organization

A multinational company migrating to Google Cloud has headquarters managing connections to offices worldwide. Each country has its own projects, but management wants integrated organization while maintaining project independence.

A. Peered VPC
B. Cloud Interconnect
C. Shared VPC
D. Cloud VPN and Cloud Router

Correct Answer: C

Shared VPC creates a single, global VPC organized by a central host project. Service projects in each country maintain independence for their resources while the network infrastructure is centrally managed. This balances control policies at the network level with freedom to manage application projects. whizlabs

VPC Peering provides connectivity but no organizational hierarchy or centralized management. Cloud Interconnect connects on-premises networks. Cloud VPN and Cloud Router provide on-premises connectivity, not inter-project networking. whizlabs

---

Section 7: AI/ML integration and solutions (New October 2025 content)

Question 28: Unified ML platform selection

A media company wants to create prediction models with minimal code, develop customized models with multiple open source frameworks, integrate teamwork for MLOps, and serve models optimally.

Which service best meets these requirements?

A. Video Intelligence API
B. TensorFlow Enterprise and KubeFlow
C. BigQuery ML
D. Vertex AI
E. Kubernetes and TensorFlow Extend

Correct Answer: D

Vertex AI is Google's unified ML platform that integrates multiple tools and MLOps pipelines. Whizlabs It exploits AutoML for minimal/no-code experimental models while supporting custom model development with TensorFlow, PyTorch, and scikit-learn. It enables continuous modeling through TensorFlow Extended and Kubeflow Pipelines integration, and provides optimized model serving with automatic scaling. whizlabs

Video Intelligence API uses pre-trained models lacking customization. TensorFlow Enterprise/KubeFlow covers only custom models and MLOps. BigQuery ML requires data in BigQuery and has limited model types. Raw Kubernetes lacks the integrated ML tooling. whizlabs

---

Question 29: Time series forecasting in BigQuery

You're developing a machine learning model to forecast daily inventory demand for a retailer. The dataset in BigQuery contains 2 years of daily records with features like product ID, region, holiday status, and promotional events. The model must capture time-based patterns like weekly seasonality and handle categorical features efficiently.

What should you do?

A. Use BigQuery ML with CREATE MODEL statement and ARIMA_PLUS
B. Preprocess data with Dataproc Spark, then train a custom LSTM model on Vertex AI
C. Use Dataflow for windowed aggregations and Vertex AI AutoML Tabular
D. Export data to Cloud Storage, train using Prophet in Cloud Functions

Correct Answer: A

BigQuery ML's ARIMA_PLUS automatically incorporates lagged variables, seasonality (weekly cycles), and holiday effects without manual feature engineering. It natively handles categorical columns via one-hot encoding. Training occurs directly in BigQuery, avoiding data movement. This provides the simplest workflow with optimal results for time series forecasting.

Custom LSTMs add unnecessary complexity for structured tabular data and require GPU resources. Dataflow adds latency and AutoML Tabular doesn't handle time series as well as ARIMA_PLUS. Prophet requires manual pipeline orchestration outside BigQuery.

Question 30: ML pipeline orchestration

A platform uses XGBoost and follows CI/CD with Docker containers. They need to classify users and update models frequently. How can you optimize frequently re-trained operations with an optimized workflow system?

A. Deploy the model on BigQuery ML
B. Use Kubeflow Pipelines to design and execute your workflow
C. Use AI Platform
D. Orchestrate activities with Google Cloud Workflows
E. Develop procedures with Pub/Sub and Cloud Run

Correct Answer: B

Kubeflow Pipelines is specifically designed for creating and deploying ML workflows using Docker containers. It uses packaged templates in Docker images running in Kubernetes environments, manages various experiments and tests, simplifies pipeline orchestration, and enables component reuse. (whizlabs) This aligns perfectly with their existing CI/CD containerized approach.

Question 31: Vertex AI Pipelines for object detection

You're developing an ML model using video frames to create bounding boxes around objects. You want to automate ingestion from Cloud Storage, training with hyperparameter tuning, and deployment to an endpoint—with minimal cluster management.

What approach should you use?

A. Use Kubeflow Pipelines on Google Kubernetes Engine
B. Use Vertex AI Pipelines with TensorFlow Extended (TFX) SDK
C. Use Vertex AI Pipelines with Kubeflow Pipelines SDK
D. Use Cloud Composer for orchestration

Correct Answer: C

Per Google's documentation, for non-TFX use cases (including image/video processing), **Vertex AI Pipelines with the Kubeflow Pipelines SDK** is recommended. The SDK allows implementing workflows with custom components or prebuilt Google Cloud Pipeline Components. Vertex AI Pipelines provides **managed, serverless execution** with minimal cluster management.

TFX is recommended for structured/text data processing at scale, not image/video workloads. Running Kubeflow directly on GKE requires cluster management overhead.

Question 32: GKE workload identity for ML services

Your company uses Kubernetes and GKE. They want an open platform without vendor lock-ins but need to securely access advanced GCP APIs using standard methodologies.

Which solution do you recommend?

A. API keys
B. Service Accounts
C. Workload Identity
D. Workload Identity Federation

Correct Answer: C

Workload Identity is Google's recommended way for GKE applications to securely access GCP APIs. It configures Kubernetes service accounts to authenticate as corresponding Google service accounts, managing identities and authorization securely while maintaining Kubernetes' open platform approach.

API keys offer minimal security without authorization capabilities. GCP Service Accounts are proprietary while Kubernetes uses its own service account system. Workload Identity Federation is for external identity providers like AWS IAM or Azure AD.

Question 33: Serverless microservices integration

A company is migrating monolithic applications into containerized RESTful microservices using Cloud Run. They want to gradually migrate, activating new microservices while maintaining the legacy application.

How can they integrate the legacy app with new microservices? (Choose 3)

A. Use an HTTP(S) Load Balancer
B. Develop a proxy inside the monolithic application
C. Use Cloud Endpoints/Apigee
D. Use Serverless NEGs for integration
E. Use App Engine Flexible Edition

Correct Answer: A, C, D

HTTP(S) Load Balancing with Serverless NEGs enables directing traffic to Cloud Run services based on URL paths. **Network endpoint groups (NEGs)** become target proxies with URL maps performing forwarding, enabling seamless integration with legacy applications. **API Management (Cloud Endpoints/Apigee)** creates a façade that integrates different applications uniformly.

Developing a proxy inside the monolith requires constant updates and service interruptions. App Engine Flexible cannot integrate legacy monolithic applications with new serverless functions.

Question 34: Fault injection testing in GKE

You're developing microservices on GKE. During testing, you want to validate application behavior if a specific microservice crashes.

What should you do?

A. Add a taint to a node and configure pod anti-affinity
B. Use Istio's fault injection on the particular microservice
C. Destroy one of the Kubernetes cluster nodes
D. Configure Istio's traffic management to steer traffic away

Correct Answer: B

Istio fault injection is a chaos engineering technique that deliberately introduces errors to test system resilience. It simulates failures (delays, HTTP errors) without actually stopping the service, validating how the rest of the application reacts. VirtualServices inject faults without updating application code.

Taints and anti-affinity don't simulate crashes. Destroying a node affects multiple microservices, not just the target. Traffic management redirects traffic but doesn't simulate crash behavior.

Question 35: Anthos Service Mesh troubleshooting

You installed Anthos Service Mesh on a GKE cluster. A microservice has increased latency when calling other microservices.

How can you troubleshoot?

A. Use the Service Mesh visualization in Cloud Console to inspect telemetry
B. Use Anthos Config Management to create a ClusterSelector
C. Use Anthos Config Management to create a namespaceSelector
D. Reinstall Istio using the default profile

Correct Answer: A

Anthos Service Mesh provides robust tracing, monitoring, and logging features for deep service performance insights. The Service Mesh pages in Cloud Console show summary metrics, charts, graphs, **service topology visualization**, and **request latency analysis** between microservices—enabling identification of upstream service latency issues.

Config Management handles policy synchronization, not performance debugging. Reinstalling Istio doesn't address the underlying latency cause.

Question 36: Multi-cloud service mesh

Your company has Kubernetes projects across GCP and AWS with many inter-relationships. You need to monitor functionality, performance, and security across this complex multi-cloud environment.

Which service helps?

A. Traffic Director
B. Istio on GKE
C. Apigee
D. App Engine Flexible Edition

Correct Answer: A

Traffic Director is a fully managed service mesh control plane for real-time monitoring, security, and telemetry data collection across **multi-cloud microservices environments** including GCP, AWS, and on-premises. It provides consistent service discovery, load balancing, and observability across all environments.

Istio on GKE only covers GKE within GCP. Apigee manages API gateway functionality, not service-to-service mesh communication. App Engine Flexible is PaaS within Google Cloud only.

Additional high-value practice questions

Question 37: Data labeling for ML

You have a classification model with lots of unlabeled data from a Data Lake and limited time.

Which service could help?

A. Vertex Data Labeling
B. Mechanical Turk
C. GitLab ML
D. Tag Manager

Correct Answer: A

Vertex AI Data Labeling Service provides professional human labelers to prepare correct labels for training data. You provide the dataset, vocabulary of possible labels, and instruction document. Labelers follow your directions to complete labeling, fully integrated with the Vertex AI workflow.

Question 38: Explaining model predictions

How can you explain model prediction results for stakeholders?

A. Use Vertex AI Workbench
B. Use Vertex AI Pipelines
C. Use Vertex AI Explainable AI

Correct Answer: C

Vertex AI Explainable AI provides feature attributions showing how much each feature contributed to predictions. It supports Integrated Gradients for deep neural networks, SHAP values for tabular data models, and feature importance visualization for understanding model behavior.

Question 39: Large data transfer from Azure

Data engineers are transferring approximately 50 TB from Microsoft Azure to Google Cloud Storage following Google recommended practices.

Which method should they use?

A. Use Storage Transfer Service
B. Copy data with gsutil
C. Copy data with bq
D. Use Transfer Appliance

Correct Answer: A

Storage Transfer Service is specifically designed for large-scale data transfers between cloud providers. It handles the 50TB transfer efficiently with automatic retry, parallel transfers, and bandwidth management. Transfer Appliance is for on-premises data. gsutil is inefficient for 50TB transfers.

Question 40: PCI DSS scope reduction

A travel booking platform wants to reduce PCI DSS scope while keeping the ability to analyze purchase behavior and payment trends.

Which approach should you adopt?

A. Export Cloud Logging to BigQuery and restrict auditor access
B. Place all components handling cardholder data in a separate project
C. Implement a tokenization service and persist only tokens
D. Create dedicated subnetworks and isolate services processing cardholder data

Correct Answer: C

Tokenization replaces primary account numbers with tokens before data enters your applications. Only the vault or payment gateway retains actual sensitive card data. Your databases, logs, and analytics pipelines handle only tokens, falling outside most PCI DSS controls while still enabling behavioral analysis.

Key exam preparation tips

Understand the Well-Architected Framework pillars: Operational Excellence, Security, Reliability, Performance Optimization, Cost Optimization, and Sustainability are now woven throughout all exam objectives.

Study the new case studies thoroughly: Altostrat Media, Cymbal Retail, EHR Healthcare, and KnightMotives Automotive all incorporate AI integration requirements. Case studies appear on split-screen during the exam and comprise 20-30% of questions.

Master Vertex AI comprehensively: New sections 2.4 and 2.5 cover Vertex AI Pipelines, AutoML, custom training, model deployment, and MLOps workflows. Understand when to use BigQuery ML versus Vertex AI.

Practice scenario-based reasoning: The exam emphasizes selecting the most appropriate solution for specific business requirements, not just knowing what services do.

Know the networking fundamentals: Shared VPC, VPC peering, Private Google Access, Serverless VPC Access connectors, and Cloud Interconnect selection criteria appear frequently.

Understand security boundaries: IAM hierarchy, Organization Policies, VPC Service Controls, and Workload Identity are essential security topics with expanded coverage.

Official resources to use: Google Cloud Skills Boost learning path (cloudskillsboost.google/paths/12), official sample questions form, and the Architecture Center (cloud.google.com/architecture) for Well-Architected Framework content.