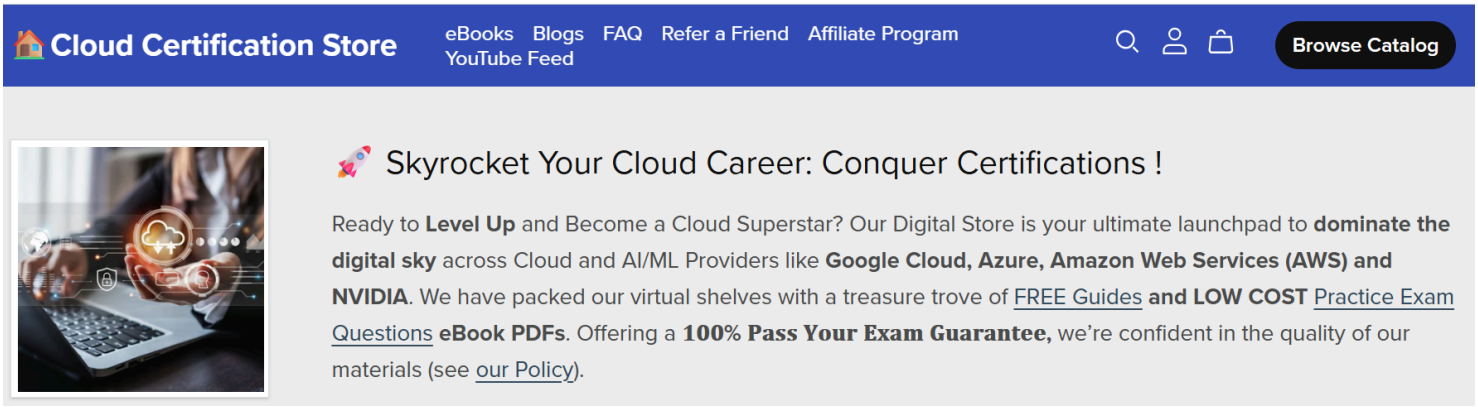
CloudCertificationStore.com



CloudCertificationStore.com is an online platform dedicated to helping professionals prepare for **cloud and AI/ML certifications** across **Google Cloud, Microsoft Azure, Amazon Web Services (AWS), and NVIDIA**.

It offers **affordable, high-quality practice exam PDFs and eBooks**—typically under **USD $10**—covering both **associate** and **professional-level certifications**. Each set includes **realistic exam-style questions**, **detailed explanations**, and **exam readiness checklists** designed to help learners **practice, assess, and pass with confidence**.

The store regularly releases new titles and bundles such as:

- 🌐 **Multi-Cloud AI & ML Bundles** (Buy 2 Get 1 Free)
- ☘️ **Google Cloud Professional Series** (Architect, Developer, Data Engineer, ML Engineer, Database Engineer, and more)
- 💙 **Microsoft Azure Certifications** (AZ-204, AZ-305, AZ-500, DP-100, and others)
- 🧡 **AWS Certification Tracks** (SAA-C03, DVA-C02, DOP-C02, CLF-C02, etc.)
- 💚 **NVIDIA Certifications** (NCP-AAI, NCA-AIIO, and related AI Infrastructure titles)

Each product is designed by experienced cloud architects and exam specialists to ensure **accuracy, clarity, and real-world relevance**—helping learners **save time, focus on weak areas, and master the actual exam format**.

🔗 Visit: https://cloudcertificationstore.com/

📚 "**Real Exam. Real Success.**"

🛠️ Introducing the Multi-Cloud Bundle: Machine Learning (ML) – Buy 2 Get 1 Free 🛠️

https://cloudcertificationstore.com/b//apkZU

🛠️ Multi-Cloud Bundle: Machine Learning (ML) – Buy 2 Get 1 Free 🛠️

https://cloudcertificationstore.com/b//apkZU

📘 About the Bundle

Machine Learning is at the heart of every modern cloud platform — and this **Multi-Cloud ML Practice Exam Bundle** brings together the three biggest ecosystems: **Google Cloud, AWS, and Microsoft Azure**.

Whether you're training models on Vertex AI, SageMaker, or Azure Machine Learning, this bundle gives you the ultimate study toolkit to become **multi-cloud ML certified** — all while saving big.

✅ What's in the Practice Exam Questions Bundle?

This bundle includes **3 full eBooks** built to simulate real exam conditions for each platform's Machine Learning certification:

AWS Certified Machine Learning Engineer – Associate (MLA-C01)

Valued at $9.95

🔗 https://cloudcertificationstore.com/b/HaeRy

Google Professional Machine Learning Engineer (GOOG-PMLE-0010)

Valued at $9.95

🔗 https://cloudcertificationstore.com/b/Y1OyW

Microsoft Certified Azure Data Scientist Associate (DP-100)

Valued at $9.95

🔗 https://cloudcertificationstore.com/b/hvask

Individually, these would cost nearly **$30**. With this bundle, you get all three for just **$19.95** — essentially a **buy 2, get 1 free (*or you will get it under $15 if there is a 50% off full price promotion*).**

🎯 Why This Bundle?

Covers **all three major cloud providers** (AWS, Azure, Google Cloud)

Focused on **end-to-end ML workflows** across AWS, Azure, and Google Cloud

📄 **1,000+** practice questions combined when you add the 3 eBooks together.

✅ Detailed explanations for correct & incorrect answers

📊 Domain-weighted coverage following each provider's official exam guide

🧠 Practical, scenario-based challenges designed by real cloud ML engineers

🧠 Final review checklists & exam readiness scorecards for each exam

One-time purchase — **download, keep forever, and study offline**

Study **offline, anytime** — one-time PDF download

💡 Who Is This Bundle For?

Cloud engineers or data scientists pursuing **multi-cloud ML fluency**

Professionals expanding from one provider (e.g., AWS → Azure ML or GCP Vertex AI)

Anyone preparing for AI/ML certifications and wanting cross-platform confidence

💬 **Learner Feedback**

“I used this bundle to compare SageMaker vs. Vertex AI vs. Azure ML workflows — perfect for understanding each platform.”

— *Victor A., Machine Learning Engineer, Brazil*

“Worth every dollar. The explanations actually teach — not just test.”

— *Sandra L., AI Practitioner, Germany*

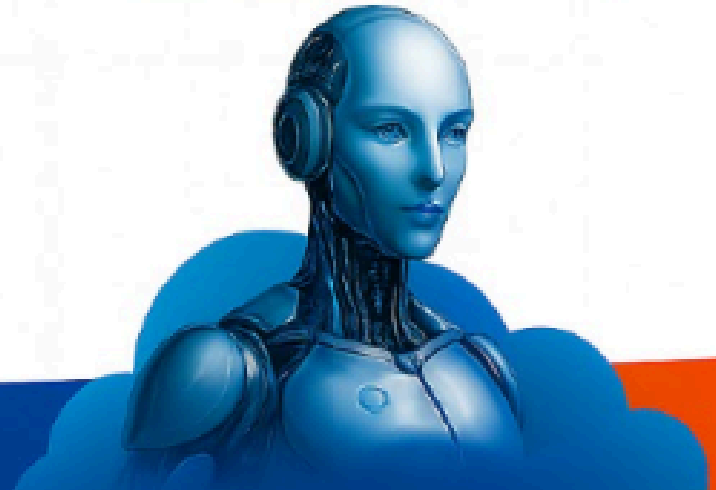
Check out below PREVIEWS for the 3 eBooks

#MachineLearning #ArtificialIntelligence #MultiCloud #GoogleCloud #AWS #MicrosoftAzure #AzureML #VertexAI #SageMaker #DataScience #MLEngineer #AIEngineer #CloudAI #CloudCertifications #CertificationGoals #ExamPrep #StudyGuide #PracticeExam #PassTheExam #CloudCertificationStore #DigitalDownload #AffordablePrep #SelfPacedLearning #AIReadiness #CareerUpgrade #CloudCareers #MLStudy #MLCertification #TechCertifications #CloudComputing #AIEngineerCareer #CloudSkills #ExamReadiness #MLOps #DataEngineer #CloudTraining #CloudLearning #MLBundle #MultiCloudEngineer #MachineLearningBundle

aws

AWS CERTIFIED MACHINE LEARNING ENGINEER

ASSOCIATE



MLA-C01

PRACTICE EXAM QUESTIONS

CloudCertification Store

AWS Certified Machine Learning Engineer Associate MLA-C01 - Practice Exam Questions (AWS-MLA-C01-0010)

DISCLAIMER

- This practice test includes questions **compiled from various exam preparation platforms.**
- Important: **Questions and answers were AI-assisted and human-curated**. Verify accuracy with official documentation before relying on this material.
- **Users are strongly encouraged to** double-check all content against **official documentation and trusted sources** before using it for exam preparation or making important decisions.
- The creators of this material assume **no responsibility** for any errors, inaccuracies, or outcomes, including exam results, based on the use of this content.
- **Some questions might be duplicated or close** to previous ones, this is done on purpose as a way to re-inforce your learning.
- Single-user licence only
    - Includes one unique Payhip Licence Key per purchase, along with a Product Key.
    - Redistribution, resale, or public posting is prohibited. We can trace any file to the purchaser, with the use of the purchased License Key and Product Key.

AWS Certified Machine Learning Engineer Associate MLA-C01 - Practice Exam Questions (AWS-MLA-C01-0010)

AWS Certified Machine Learning Engineer – Associate

Issued by Amazon Web Services Training and Certification

Earners of this badge have knowledge and skills in developing, deploying, maintaining, and monitoring ML solutions to meet AI/ML objectives. They know how to ingest, transform, validate, and prepare data for ML modeling. They have skills in implementing and operationalizing ML workloads in production. They can select modeling approaches and analyze model performance. They have the expertise to monitor ML solutions and to secure ML systems and resources.

https://aws.amazon.com/certification/certified-machine-learning-engineer-associate/

Exam overview

AWS Certified Machine Learning Engineer - Associate

Category

Associate

Exam duration

130 minutes

Exam format

65 questions

Cost

150 USD. Visit Exam pricing for additional cost information, including foreign exchange rates

Intended candidate

Individuals with at least 1 year of experience using Amazon SageMaker and other ML engineering AWS services

Candidate role examples

Backend software developer, DevOps engineer, data engineer, MLOps engineer, and data scientist

Testing options

Pearson VUE testing center or online proctored exam

Prepare for the exam

Go from start to certified. Follow our Exam Prep Plan on AWS Skill Builder, our online learning center, so you can approach exam day with confidence.

1. Get to know the exam with exam-style questions

Follow the 4-step plan.

Review the exam guide.

Take the AWS Certification Official Practice Question Set to understand exam-style questions.

Take the AWS Certification Official Pretest to identify any areas where you need to refresh your AWS knowledge and skills.

2. Refresh your AWS knowledge and skills

Enroll in digital courses where you need to fill gaps in knowledge and skills, practice with AWS Builder Labs, AWS Cloud Quest, and AWS Jam.

3. Review and practice for your exam

Review the scope of the exam. Explore each exam domain's topics and how they align to AWS services. Reinforce your knowledge and identify learning gaps with exam-style questions and flashcards. Follow instructors as they walk through exam-style questions and provide test-taking strategies. Continue practicing with AWS Builder Labs and/or AWS SimuLearn.

4. Assess your exam readiness

Take the AWS Certification Official Practice Exam.

Key FAQs to help you get started

Who should earn AWS Certified Machine Learning Engineer - Associate?

The ideal candidate for this exam has at least 1 year of experience in machine learning engineering or a related field and 1 year of hands-on experience with AWS services. Professionals who do not have prior machine learning experience can take the training available in the Exam Prep Plans and get started building their knowledge and skills.

How will the AWS Certified Machine Learning Engineer - Associate help my career?

Per the World Economic Forum Future of Jobs Report 2023, demand for AI and Machine Learning Specialists is expected to grow by 40%. However, 70% of North American IT leaders say they have the greatest difficulty filling AI/ML specialist roles. This certification can position you for in-demand machine learning jobs in AWS Cloud.

How is AWS Certified Machine Learning Engineer - Associate different from AWS Certified Machine Learning - Specialty?

AWS Certified Machine Learning Engineer - Associate is a role-based certification designed for ML engineers and MLOps engineers with at least one year of experience in AI/ML.

AWS Machine Learning - Specialty is a specialty certification covering topics across data engineering, data analysis, modeling, and ML implementation and ops. It is more suitable for individuals with 2 or more years of experience developing, architecting, and running ML workloads on AWS.

What certification(s) should I earn next after AWS Certified Machine Learning Engineer - Associate?

For professionals looking to dive deeper into machine learning, we recommend AWS Certified Machine Learning - Specialty.

How long is this certification valid for?

This certification is valid for 3 years. Before your certification expires, you can recertify by passing the latest version of this exam. Learn more about recertification options for AWS Certifications.

Additional resources

Learn more about AWS Certification exams

Before scheduling your AWS Certification exam, review the available options for the specific exam and your desired exam language.

View all exams

AWS Training Live on Twitch

Access free, live, and on-demand training on our dedicated Twitch channel. Join AWS experts for live shows, chat with the community, or explore on-demand training.

Explore more

AWS Certification FAQs

Questions about AWS Certification? Browse frequently asked questions about getting AWS Certified and AWS Certification.

Browse AWS Certification FAQ

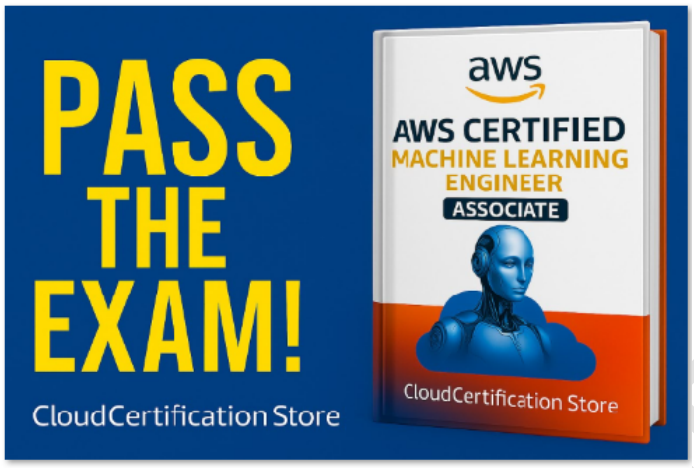
Information and policies

Learn about what to expect with scheduling your exam, identification requirements, exam functionality, relevant policies, and more.
Explore more

Exam vouchers

Are you supporting a team taking AWS Certification exams? Make it easier with exam vouchers. Purchase online anytime, and then use a self-service portal to efficiently distribute, track, and manage standard exam vouchers.

Get exam vouchers



Practice Questions

Question 1

An ML engineer needs to use data with Amazon SageMaker Canvas to train an ML model. The data is stored in Amazon S3 and is complex in structure. The ML engineer must use a file format that minimizes processing time for the data.

Which file format will meet these requirements?

A. CSV files compressed with Snappy

B. JSON objects in JSONL format

C. JSON files compressed with gzip

D. Apache Parquet files

✅ **Correct answer: D. Apache Parquet files**

Parquet is a columnar, compressed, and splittable format that minimizes I/O and speeds up reads for analytical workloads, which is ideal for complex, nested structures. Canvas and downstream AWS analytics services optimize for Parquet, reducing the time spent parsing and transforming records. Column pruning and predicate pushdown further reduce processing time when only a subset of

columns is needed. This combination yields faster data ingest and transformation than row-oriented formats like CSV or JSON.

Incorrect answers:

❌ **A. CSV files compressed with Snappy** – CSV is a row-oriented, text-based format that lacks schema information and efficient columnar access, so it incurs high parsing overhead. While Snappy helps with compression/decompression speed, it does not fix CSV's lack of type fidelity nor enable predicate pushdown. CSV also struggles with complex or nested data, requiring additional parsing logic. Overall, it results in more CPU time and slower ingestion for complex datasets.

❌ **B. JSON objects in JSONL format** – JSONL preserves one JSON object per line, which is convenient for streaming but still row-oriented and verbose. Complex nested structures are supported, but the lack of columnar layout means no column pruning and heavier parsing costs. Compression is not inherent and, if applied, increases CPU time for decompression. As a result, JSONL typically increases processing time compared to Parquet.

❌ **C. JSON files compressed with gzip** – Gzip compression reduces storage, but decompression is CPU-intensive and not splittable by default, hindering parallelism on large files. JSON is row-oriented and verbose, further increasing parse time for complex structures. There is no native schema enforcement, which

can add overhead in downstream processing. Together these factors make gzip-compressed JSON slower than Parquet for Canvas workflows.

Question 2

A company is using an AWS Lambda function to monitor the metrics from an ML model. An ML engineer needs to implement a solution to send an email message when the metrics breach a threshold.

Which solution will meet this requirement?

A. Log the metrics from the Lambda function to AWS CloudTrail. Configure a CloudTrail trail to send the email message.

B. Log the metrics from the Lambda function to Amazon CloudFront. Configure an Amazon CloudWatch alarm to send the email message.

C. Log the metrics from the Lambda function to Amazon CloudWatch. Configure a CloudWatch alarm to send the email message.

D. Log the metrics from the Lambda function to Amazon CloudWatch. Configure an Amazon CloudFront rule to send the email message.

✅ Correct answer: C. Log the metrics from the Lambda function to Amazon CloudWatch. Configure a CloudWatch alarm to send the email message.

CloudWatch is the native destination for Lambda metrics and custom application metrics, enabling immediate alarm creation. CloudWatch Alarms integrate with Amazon SNS to send email notifications when thresholds are breached. This approach is fully managed and requires minimal code: emit metrics and define an alarm with an SNS subscription. It also provides dashboards and history for easy operational visibility.

Incorrect answers:

❌ **A. Log the metrics from the Lambda function to AWS CloudTrail. Configure a CloudTrail trail to send the email message.** – CloudTrail records control-plane API calls, not arbitrary application metrics. It is not designed for metric thresholds or real-time metric evaluation. You cannot create metric alarms directly from CloudTrail events for this use case. This would be a misuse of the service and add unnecessary complexity.

❌ **B. Log the metrics from the Lambda function to Amazon CloudFront. Configure an Amazon CloudWatch alarm to send the email message.** – CloudFront is a content delivery network and does not serve as a general-purpose metrics sink. While CloudFront has its own metrics, it's unrelated to Lambda application metrics. Routing metrics through CloudFront adds no value and

introduces confusion. CloudWatch remains the correct place for metrics and alarms.

❌ **D. Log the metrics from the Lambda function to Amazon CloudWatch. Configure an Amazon CloudFront rule to send the email message.** – Even with metrics in CloudWatch, CloudFront has no role in sending notifications for metric breaches. CloudWatch Alarms with SNS subscriptions are the supported mechanism for email alerts. Introducing CloudFront offers no benefit and complicates the design. The simplest, standard path is CloudWatch → Alarm → SNS (email).

Question 3

An ML engineer needs to implement a solution to host a trained ML model. The rate of requests to the model will be inconsistent throughout the day.

The ML engineer needs a scalable solution that minimizes costs when the model is not in use. The solution also must maintain the model's capacity to respond to requests during times of peak usage.

Which solution will meet these requirements?

A. Create AWS Lambda functions that have fixed concurrency to host the model. Configure the Lambda functions to automatically scale based on the number of requests to the model.

B. Deploy the model on an Amazon Elastic Container Service (Amazon ECS) cluster that uses AWS Fargate. Set a static number of tasks to handle requests during times of peak usage.

C. Deploy the model to an Amazon SageMaker endpoint. Deploy multiple copies of the model to the endpoint. Create an Application Load Balancer to route traffic between the different copies of the model at the endpoint.

D. Deploy the model to an Amazon SageMaker endpoint. Create SageMaker endpoint auto scaling policies that are based on Amazon CloudWatch metrics to adjust the number of instances dynamically.

✅ **Correct answer: D. Deploy the model to an Amazon SageMaker endpoint. Create SageMaker endpoint auto scaling policies that are based on Amazon CloudWatch metrics to adjust the number of instances dynamically.**

SageMaker real-time endpoints support automatic scaling based on utilization metrics (e.g., invocations per instance), preserving performance under load and reducing capacity when idle. This directly addresses fluctuating traffic while minimizing cost. It also avoids operational overhead by letting the service manage

scaling rather than maintaining custom logic. The approach remains compatible with production features like Model Monitor and multi-variant testing.

Incorrect answers:

❌ **A. Create AWS Lambda functions that have fixed concurrency to host the model. Configure the Lambda functions to automatically scale based on the number of requests to the model.** – Lambda is not ideal for stateful model servers or large containerized ML runtimes, and "fixed concurrency" contradicts the need to scale elastically. Packaging heavy frameworks into Lambda often leads to cold start challenges and deployment complexity. You would still need to build custom monitoring and autoscaling behavior. SageMaker provides a purpose-built, managed inference platform.

❌ **B. Deploy the model on an Amazon Elastic Container Service (Amazon ECS) cluster that uses AWS Fargate. Set a static number of tasks to handle requests during times of peak usage.** – A static task count wastes resources during off-peak periods and fails to minimize costs when traffic is low. While ECS with Fargate can scale, the option explicitly fixes capacity, which violates the requirement. You would also need to build additional observability and autoscaling policies manually. SageMaker's integrated scaling is simpler and more targeted.

❌ **C. Deploy the model to an Amazon SageMaker endpoint. Deploy multiple copies of the model to the endpoint. Create an Application Load Balancer to route traffic between the different copies of the model at the endpoint.** – SageMaker endpoints already handle load balancing and scaling, so adding an ALB is redundant and adds complexity. Manually managing "copies" does not guarantee elasticity or cost efficiency. Autoscaling in SageMaker is the supported pattern for demand variability. This answer increases operational overhead without benefits.

Question 4

A company has a binary classification model in production. An ML engineer needs to develop a new version of the model.

The new model version must maximize correct predictions of positive labels and negative labels. The ML engineer must use a metric to recalibrate the model to meet these requirements.

Which metric should the ML engineer use for the model recalibration?

A. Accuracy

B. Precision

C. Recall

D. Specificity

✅ **Correct answer: A. Accuracy**

Accuracy directly measures the proportion of all correct predictions, combining both true positives and true negatives. If the goal is to maximize correctness across both classes, optimizing thresholding by accuracy aligns with that objective. While not always ideal for imbalanced datasets, the prompt explicitly emphasizes correct positives and negatives together. Thus, accuracy serves as the most straightforward recalibration metric here.

Incorrect answers:

❌ **B. Precision** – Precision focuses on the quality of positive predictions (TP / (TP + FP)) and ignores how well negatives are identified. Optimizing solely for precision could reduce false positives at the expense of missing many true positives. This does not align with maximizing correctness across both classes. It provides a partial view rather than a holistic measure.

❌ **C. Recall** – Recall (TP / (TP + FN)) emphasizes capturing as many positives as possible, potentially increasing false positives. Maximizing recall may degrade performance on negative classifications, hurting overall correctness. If both

positive and negative correctness matter equally, recall alone is insufficient. It's a targeted metric, not a balanced one.

❌ **D. Specificity** – Specificity (TN / (TN + FP)) focuses only on correctly identifying negatives. While valuable in certain domains, it ignores positive classification performance. Maximizing specificity might severely lower recall, which undermines the "maximize correct positives and negatives" requirement. It is not an all-encompassing correctness metric like accuracy.

Question 5

An ML engineer needs to deploy ML models to get inferences from large datasets in an asynchronous manner. The ML engineer also needs to implement scheduled monitoring of the data quality of the models. The ML engineer must receive alerts when changes in data quality occur.

Which solution will meet these requirements?

A. Deploy the models by using scheduled AWS Glue jobs. Use Amazon CloudWatch alarms to monitor the data quality and to send alerts.

B. Deploy the models by using scheduled AWS Batch jobs. Use AWS CloudTrail to monitor the data quality and to send alerts.

C. Deploy the models by using Amazon Elastic Container Service (Amazon ECS) on AWS Fargate. Use Amazon EventBridge to monitor the data quality and to send alerts.

D. Deploy the models by using Amazon SageMaker batch transform. Use SageMaker Model Monitor to monitor the data quality and to send alerts.

✅ **Correct answer: D. Deploy the models by using Amazon SageMaker batch transform. Use SageMaker Model Monitor to monitor the data quality and to send alerts.**

Batch transform is designed for asynchronous, large-scale inference without managing endpoint uptime. SageMaker Model Monitor natively profiles data quality on a schedule and integrates with CloudWatch/SNS for alerts. This pair provides an end-to-end, managed solution with minimal operational overhead. It avoids building custom pipelines for both inference scheduling and data drift detection.

Incorrect answers:

❌ **A. Deploy the models by using scheduled AWS Glue jobs. Use Amazon CloudWatch alarms to monitor the data quality and to send alerts.** – Glue is an ETL service, not an ML inference service, so you would be crafting nonstandard inference logic within ETL jobs. CloudWatch alarms require you to define and

emit your own quality metrics, increasing custom work. There is no built-in ML data quality profiling. This adds complexity compared to Model Monitor's native capabilities.

❌ **B. Deploy the models by using scheduled AWS Batch jobs. Use AWS CloudTrail to monitor the data quality and to send alerts.** – While Batch can orchestrate containers, it lacks ML-specific features for data quality. CloudTrail tracks API calls, not dataset characteristics or statistical drift, so it's unsuitable for data quality monitoring. You would need to implement comprehensive custom checks and alerting. This is higher effort and less maintainable.

❌ **C. Deploy the models by using Amazon Elastic Container Service (Amazon ECS) on AWS Fargate. Use Amazon EventBridge to monitor the data quality and to send alerts.** – ECS can run inference containers, but EventBridge is an event router, not a data quality analyzer. You would still need to compute drift statistics, thresholds, and alerting logic. This shifts significant engineering burden onto your team. SageMaker's managed services are purpose-built for these needs.

Question 6

A company has an ML model that needs to run one time each night to predict stock values. The model input is 3 MB of data that is collected during the current day. The model produces the predictions for the next day. The prediction process takes less than 1 minute to finish running.

How should the company deploy the model on Amazon SageMaker to meet these requirements?

A. Use a multi-model serverless endpoint. Enable caching.

B. Use an asynchronous inference endpoint. Set the InitialInstanceCount parameter to 0.

C. Use a real-time endpoint. Configure an auto scaling policy to scale the model to 0 when the model is not in use.

D. Use a serverless inference endpoint. Set the MaxConcurrency parameter to 1.

✅ **Correct answer: D. Use a serverless inference endpoint. Set the MaxConcurrency parameter to 1.**

Serverless inference removes idle capacity cost when the endpoint is not invoked, which suits a once-per-night job. With a very short runtime and tiny input, a minimal concurrency setting is sufficient and cost-efficient. You avoid managing instances or scaling policies entirely. Cold starts are negligible here given the single, scheduled invocation window.

Incorrect answers:

❌ **A. Use a multi-model serverless endpoint. Enable caching.** – Multi-model endpoints are beneficial when hosting many models behind one endpoint, which is not required here. Caching does not materially improve a once-per-day, sub-minute job with small input. This adds unnecessary complexity without cost or performance benefits. A single serverless endpoint is simpler and cheaper.

❌ **B. Use an asynchronous inference endpoint. Set the InitialInstanceCount parameter to 0.** – Asynchronous endpoints are designed for larger payloads or long-running requests with queueing, which is overkill for a 1-minute nightly job. Also, asynchronous endpoints do not use "InitialInstanceCount" in the way real-time variants do. Serverless provides the simplest pay-per-invocation model. This option adds operational features you don't need.

❌ **C. Use a real-time endpoint. Configure an auto scaling policy to scale the model to 0 when the model is not in use.** – Real-time endpoints do not natively scale down to zero; you would pay for idle capacity. Managing lifecycle (start/stop) to emulate zero is additional overhead. For a nightly run, continuous provisioning is wasteful. Serverless avoids these costs and management tasks.

Question 7

An advertising company uses AWS Lake Formation to manage a data lake. The data lake contains structured data and unstructured data. The company's ML engineers are assigned to specific advertisement campaigns.

The ML engineers must interact with the data through Amazon Athena and by browsing the data directly in an Amazon S3 bucket. The ML engineers must have access to only the resources that are specific to their assigned advertisement campaigns.

Which solution will meet these requirements in the MOST operationally efficient way?

A. Configure IAM policies on an AWS Glue Data Catalog to restrict access to Athena based on the ML engineers' campaigns.

B. Store users and campaign information in an Amazon DynamoDB table. Configure DynamoDB Streams to invoke an AWS Lambda function to update S3 bucket policies.

C. Use Lake Formation to authorize AWS Glue to access the S3 bucket. Configure Lake Formation tags to map ML engineers to their campaigns.

D. Configure S3 bucket policies to restrict access to the S3 bucket based on the ML engineers' campaigns.

✅ **Correct answer: C. Use Lake Formation to authorize AWS Glue to access the S3 bucket. Configure Lake Formation tags to map ML engineers to their campaigns.**

Lake Formation provides fine-grained, tag-based access control across both cataloged tables (for Athena) and underlying S3 data locations. By using LF-tags tied to campaigns, you centralize and scale permissions management without handcrafting policies per user or dataset. This keeps governance consistent for both query and direct S3 access. It's the native, operationally efficient path for campaign-scoped permissions in a data lake.

Incorrect answers:

❌ **A. Configure IAM policies on an AWS Glue Data Catalog to restrict access to Athena based on the ML engineers' campaigns.** – IAM alone does not easily express fine-grained, column/table-level entitlements across evolving datasets. It also doesn't propagate smoothly to S3 object-level controls for the same logical datasets. Maintaining bespoke IAM for every campaign increases complexity. Lake Formation's tag-based model is purpose-built for this.

❌ **B. Store users and campaign information in an Amazon DynamoDB table. Configure DynamoDB Streams to invoke an AWS Lambda function to update S3 bucket policies.** – This creates a custom entitlement system you must build,

test, and maintain. It doesn't natively integrate with Athena's view of data, risking policy drift between S3 and the catalog. Automated policy updates add operational fragility. Lake Formation already solves this use case without custom code.

❌ **D. Configure S3 bucket policies to restrict access to the S3 bucket based on the ML engineers' campaigns.** – S3 bucket policies alone are coarse and hard to maintain for campaign-level granularity across many prefixes/objects. You also need to align permissions with Athena's Data Catalog to avoid mismatches. This approach quickly becomes unmanageable at scale. Lake Formation unifies governance across query and storage layers.

Question 8

A company's ML engineer has deployed an ML model for sentiment analysis to an Amazon SageMaker endpoint. The ML engineer needs to explain to company stakeholders how the model makes predictions.

Which solution will provide an explanation for the model's predictions?

A. Use SageMaker Model Monitor on the deployed model.

B. Use SageMaker Clarify on the deployed model.

C. Show the distribution of inferences from A/B testing in Amazon CloudWatch.

D. Add a shadow endpoint. Analyze prediction differences on samples.

✅ **Correct answer: B. Use SageMaker Clarify on the deployed model.**

SageMaker Clarify provides model explainability reports, including SHAP-based feature attributions that show how inputs influence predictions. It integrates directly with endpoints to generate explanations on production models. These insights are designed for stakeholders to understand and trust model behavior. Clarify also supports bias analysis, enhancing governance and transparency.

Incorrect answers:

❌ **A. Use SageMaker Model Monitor on the deployed model.** – Model Monitor focuses on data quality and drift, not per-prediction explainability. It won't tell stakeholders which features drove a specific outcome. While useful for operations, it doesn't answer "why" a particular prediction occurred. Clarify is the correct tool for explanations.

❌ **C. Show the distribution of inferences from A/B testing in Amazon CloudWatch.** – A/B distributions provide performance comparisons, not feature attributions. Stakeholders still won't know how inputs affected a single

prediction. This approach lacks transparency into model internals. It's complementary for experiments, not for explainability.

❌ **D. Add a shadow endpoint. Analyze prediction differences on samples.** – Shadow testing compares models under real traffic but does not expose how a model reasons. You would only see output divergences, not the feature contributions. It's an evaluation tactic, not an explainability method. Clarify directly addresses the requirement.

Question 9

A company wants to reduce the cost of its containerized ML applications. The applications use ML models that run on Amazon EC2 instances, AWS Lambda functions, and an Amazon Elastic Container Service (Amazon ECS) cluster. The EC2 workloads and ECS workloads use Amazon Elastic Block Store (Amazon EBS) volumes to save predictions and artifacts.

An ML engineer must identify resources that are being used inefficiently. The ML engineer also must generate recommendations to reduce the cost of these resources.

Which solution will meet these requirements with the LEAST development effort?

A. Create code to evaluate each instance's memory and compute usage.

B. Add cost allocation tags to the resources. Activate the tags in AWS Billing and Cost Management.

C. Check AWS CloudTrail event history for the creation of the resources.

D. Run AWS Compute Optimizer.

✅ **Correct answer: D. Run AWS Compute Optimizer.**

Compute Optimizer analyzes utilization metrics across EC2, Lambda, EBS volumes, and ECS on Fargate to surface right-sizing and configuration recommendations. It provides actionable guidance (e.g., instance families, volume types/sizes) without writing custom analysis code. This directly targets cost inefficiencies with minimal setup. The recommendations help you realize savings quickly across multiple compute modalities.

Incorrect answers:

❌ **A. Create code to evaluate each instance's memory and compute usage.** – Building a custom telemetry and analysis pipeline is time-consuming and error-prone. You would need to ingest CloudWatch metrics, implement heuristics, and maintain dashboards. This duplicates capabilities that Compute Optimizer already provides. It raises operational overhead for limited benefit.

❌ **B. Add cost allocation tags to the resources. Activate the tags in AWS Billing and Cost Management.** – Tags improve cost visibility and chargeback but do not generate optimization recommendations. You would still need to analyze usage manually to find inefficiencies. While tagging is a good practice, it does not satisfy the requirement to identify and recommend right-sizing actions. Compute Optimizer addresses that gap directly.

❌ **C. Check AWS CloudTrail event history for the creation of the resources.** – CloudTrail reveals provision events, not ongoing utilization or cost efficiency. Creation history doesn't indicate whether a resource is oversized or underutilized. This provides little to no guidance for optimization. It's not the right tool for cost reduction analysis.

Question 10

A company has deployed an ML model that detects fraudulent credit card transactions in real time in a banking application. The model uses Amazon SageMaker Asynchronous Inference. Consumers are reporting delays in receiving the inference results.

An ML engineer needs to implement a solution to improve the inference performance. The solution also must provide a notification when a deviation in model quality occurs.

Which solution will meet these requirements?

A. Use SageMaker real-time inference for inference. Use SageMaker Model Monitor for notifications about model quality.

B. Use SageMaker batch transform for inference. Use SageMaker Model Monitor for notifications about model quality.

C. Use SageMaker Serverless Inference for inference. Use SageMaker Inference Recommender for notifications about model quality.

D. Keep using SageMaker Asynchronous Inference for inference. Use SageMaker Inference Recommender for notifications about model quality.

✅ **Correct answer: A. Use SageMaker real-time inference for inference. Use SageMaker Model Monitor for notifications about model quality.**

Real-time endpoints deliver low-latency predictions suitable for interactive fraud detection and eliminate queue-induced lag from asynchronous processing. Model Monitor can continuously evaluate production data quality and emit alerts when drift or schema violations occur. This pairing directly addresses both latency and

quality monitoring requirements. It also avoids building custom alerting or data checks.

Incorrect answers:

❌ **B. Use SageMaker batch transform for inference. Use SageMaker Model Monitor for notifications about model quality.** – Batch transform is designed for offline, large-scale jobs and cannot meet interactive latency demands. It would increase, not decrease, response time for consumers. While Model Monitor is still valid, the inference mode is inappropriate. Real-time inference is the proper fit for immediate decisions.

❌ **C. Use SageMaker Serverless Inference for inference. Use SageMaker Inference Recommender for notifications about model quality.** – Serverless inference may help occasionally, but it can introduce cold starts and is not guaranteed to resolve latency concerns under bursty real-time loads. Inference Recommender evaluates configuration/performance, not model quality drift or data issues. This combination does not ensure low latency and the right alerting. Real-time endpoints plus Model Monitor is the targeted solution.

❌ **D. Keep using SageMaker Asynchronous Inference for inference. Use SageMaker Inference Recommender for notifications about model quality.** – Asynchronous inference is the cause of delays due to queueing semantics and

background processing. Inference Recommender won't notify on data quality deviations and is not intended for production monitoring. This option fails to address both the latency and the monitoring needs. Switching to real-time and enabling Model Monitor resolves the core issues.

Question 11

An ML engineer needs to use AWS CloudFormation to create an ML model that an Amazon SageMaker endpoint will host.

Which resource should the ML engineer declare in the CloudFormation template to meet this requirement?

A. AWS::SageMaker::Model

B. AWS::SageMaker::Endpoint

C. AWS::SageMaker::NotebookInstance

D. AWS::SageMaker::Pipeline

✅ **Correct answer: A. AWS::SageMaker::Model**

This resource defines the model artifacts and the container image configuration that SageMaker endpoints consume. In SageMaker, you first create a Model resource, then attach it to a hosted endpoint configuration and endpoint for

inference. Declaring the model in CloudFormation enables repeatable, version-controlled deployments. Without the Model resource, the endpoint would have nothing to serve.

Incorrect answers:

❌ **B. AWS::SageMaker::Endpoint** – An endpoint is the hosting resource but it references a Model (via an EndpointConfig). Creating only an endpoint without a defined Model does not specify the artifacts or container. Endpoints are typically created after a Model and EndpointConfig exist. Alone, it cannot satisfy the "create an ML model" requirement.

❌ **C. AWS::SageMaker::NotebookInstance** – Notebooks are for interactive development, not for defining hosted model artifacts. Spinning up a notebook does not deploy a model to an endpoint or register model containers. It provides compute for exploration rather than production inference resources.

❌ **D. AWS::SageMaker::Pipeline** – A Pipeline orchestrates ML steps (processing, training, model registration), but it does not directly define the runtime model object an endpoint needs. You still end up creating a Model resource as a step output. Pipelines alone cannot host inference without a Model declaration.

Question 12

An ML engineer is evaluating several ML models and must choose one model to use in production. The cost of false negative predictions by the models is much higher than the cost of false positive predictions.

Which metric finding should the ML engineer prioritize the MOST when choosing the model?

A. Low precision

B. High precision

C. Low recall

D. High recall

✅ **Correct answer: D. High recall**

High recall means the model captures the majority of actual positives, minimizing false negatives. When missing a positive case is much costlier than raising extra alerts, recall is the appropriate optimization target. Prioritizing recall reduces the risk of overlooking true positive events. Thresholds can later be tuned to balance precision within acceptable business limits.

Incorrect answers:

❌ **A. Low precision** – Low precision implies many false positives, which may be acceptable only if recall is very high, but "low precision" is not a desirable property. It increases unnecessary follow-up actions and noise for downstream teams. It's not a metric you would intentionally prioritize as "low." The problem statement emphasizes reducing false negatives, not encouraging false positives outright.

❌ **B. High precision** – High precision minimizes false positives, which is secondary in this scenario. Optimizing precision typically comes at the cost of recall, potentially missing true positives. That trade-off contradicts the requirement to avoid false negatives. Precision can be addressed after recall targets are met.

❌ **C. Low recall** – Low recall explicitly means many false negatives, which is the opposite of the goal. A model with low recall would routinely miss positive cases. This would create unacceptable business risk. Such a model should be deprioritized for this use case.

Question 13

A company is using an Amazon Redshift database as its single data source. Some of the data is sensitive.

A data scientist needs to use some of the sensitive data from the database. An ML engineer must give the data scientist access to the data without transforming the source data and without storing anonymized data in the database.

Which solution will meet these requirements with the LEAST implementation effort?

A. Configure dynamic data masking policies to control how sensitive data is shared with the data scientist at query time.

B. Create a materialized view with masking logic on top of the database. Grant the necessary read permissions to the data scientist.

C. Unload the Amazon Redshift data to Amazon S3. Use Amazon Athena to create schema-on-read with masking logic. Share the view with the data scientist.

D. Unload the Amazon Redshift data to Amazon S3. Create an AWS Glue job to anonymize the data. Share the dataset with the data scientist.

✅ **Correct answer: A. Configure dynamic data masking policies to control how sensitive data is shared with the data scientist at query time.**

Dynamic data masking allows policy-based redaction at query time, avoiding data movement and additional storage. It’s purpose-built to share sensitive data

responsibly with minimal engineering overhead. The source remains authoritative, and policies can be adjusted centrally without duplicating data. This meets both “no transform” and “no stored anonymized copy” constraints efficiently.

Incorrect answers:

❌ **B. Create a materialized view with masking logic** – Materialized views create and store derived data, which adds management overhead and may not align with “no storing anonymized data.” They also introduce refresh considerations and latency between source and view. This is heavier operationally than dynamic masking. It’s not the least-effort approach.

❌ **C. Unload to S3 and use Athena** – Exporting data and building schema-on-read masking in Athena breaks single-source authority and creates data sprawl. It also adds pipelines, permissions, and lifecycle considerations. This contradicts the desire to avoid transforming or storing anonymized data elsewhere. It’s more complex and riskier than in-database masking.

❌ **D. Unload to S3 and anonymize with Glue** – This explicitly stores an anonymized copy, violating the requirement. It introduces ETL jobs, code, and operational maintenance. Data drift between copies becomes a concern. It’s the opposite of a “least effort” control.

Question 14

A company has AWS Glue data processing jobs that are orchestrated by an AWS Glue workflow. The AWS Glue jobs can run on a schedule or can be launched manually.

The company is developing pipelines in Amazon SageMaker Pipelines for ML model development. The pipelines will use the output of the AWS Glue jobs during the data processing phase of model development. An ML engineer needs to implement a solution that integrates the AWS Glue jobs with the pipelines.

Which solution will meet these requirements with the LEAST operational overhead?

A. Use AWS Step Functions for orchestration of the pipelines and the AWS Glue jobs.

B. Use processing steps in SageMaker Pipelines. Configure inputs that point to the Amazon R esource Names (ARNs) of the AWS Glue jobs.

C. Use Callback steps in SageMaker Pipelines to start the AWS Glue workflow and to stop the pipelines until the AWS Glue jobs finish running.

D. Use Amazon EventBridge to invoke the pipelines and the AWS Glue jobs in the desired order.

✅ **Correct answer: C. Use Callback steps in SageMaker Pipelines to start the AWS Glue workflow and to stop the pipelines until the AWS Glue jobs finish running.**

Callback steps are designed to integrate external systems into a pipeline while preserving pipeline state and dependencies. They let you trigger a Glue workflow and block until an external completion signal arrives, keeping orchestration inside SageMaker. This yields tighter control, simpler monitoring, and less moving parts. It avoids building and maintaining a separate orchestrator.

Incorrect answers:

❌ **A. Use AWS Step Functions** – While Step Functions can orchestrate both services, it introduces another control plane and duplicate orchestration logic. You would manage state transitions across two systems, which increases complexity. The question asks for least operational overhead. Callback steps keep it native to Pipelines.

❌ **B. Use processing steps with Glue job ARNs** – Processing steps execute containers managed by SageMaker, not Glue jobs, and cannot directly "point at" Glue jobs to run them. This misunderstands the purpose of a processing step. You would still need a mechanism to trigger and await Glue completion. It's not a valid integration.

❌ **D. Use Amazon EventBridge to invoke both** – EventBridge can fan-out triggers, but coordinating run order and completion semantics becomes harder. You'd have to build custom patterns and idempotency checks. It's looser coupling and higher effort than a single pipeline with a callback step.

Question 15

A company stores time-series data about user clicks in an Amazon S3 bucket. The raw data consists of millions of rows of user activity every day. ML engineers access the data to develop their ML models.

The ML engineers need to generate daily reports and analyze click trends over the past 3 days by using Amazon Athena. The company must retain the data for 30 days before archiving the data.

Which solution will provide the HIGHEST performance for data retrieval?

A. Keep all the time-series data without partitioning in the S3 bucket. Manually move data that is older than 30 days to separate S3 buckets.

B. Create AWS Lambda functions to copy the time-series data into separate S3 buckets. Apply S3 Lifecycle policies to archive data that is older than 30 days to S3 Glacier Flexible R etrieval.

C. Organize the time-series data into partitions by date prefix in the S3 bucket.

Apply S3 Lifecycle policies to archive partitions that are older than 30 days to S3 Glacier Flexible R etrieval.

D. Put each day's time-series data into its own S3 bucket. Use S3 Lifecycle policies to archive S3 buckets that hold data that is older than 30 days to S3 Glacier Flexible R etrieval.

✅ **Correct answer: C. Organize the time-series data into partitions by date prefix in the S3 bucket. Apply S3 Lifecycle policies to archive partitions that are older than 30 days to S3 Glacier Flexible R etrieval.**

Athena performance benefits greatly from partition pruning, which reduces scanned data to just the partitions of interest (e.g., last 3 days). Using date prefixes enables efficient queries and minimal I/O. Lifecycle policies meet the 30-day retention requirement automatically. This aligns storage optimization with query performance.

Incorrect answers:

❌ **A. Keep all data without partitioning** – Without partitions, Athena scans far more data than necessary, increasing latency and cost. Manual movement of data is error-prone and operationally heavy. It provides no query-time optimization. It fails both performance and operations goals.

❌ **B. Lambda copying to separate buckets** – Creating many buckets is unnecessary and complicates governance, access, and cataloging. Copying data increases cost and introduces synchronization risks. Lifecycle and partitioning can be achieved within a single bucket. This is more complex than needed.

❌ **D. One bucket per day** – Excessive bucket sprawl is an anti-pattern and complicates permissions, listings, and metadata management. Athena also expects a coherent partitioned layout within a table rather than table-per-bucket. Lifecycle at bucket granularity is inflexible and cumbersome.

Question 16

A company is using Amazon SageMaker to create ML models. The company's data scientists need fine-grained control of the ML workflows that they orchestrate. The data scientists also need the ability to visualize SageMaker jobs and workflows as a directed acyclic graph (DAG). The data scientists must keep a running history of model discovery experiments and must establish model governance for auditing and compliance verifications.

Which solution will meet these requirements?

A. Use AWS CodePipeline and its integration with SageMaker S tudio to manage the entire ML workflows. Use SageMaker ML Lineage T racking for the running history of experiments and for auditing and compliance verifications.

B. Use AWS CodePipeline and its integration with SageMaker Experiments to manage the entire ML workflows. Use SageMaker Experiments for the running history of experiments and for auditing and compliance verifications.

C. Use SageMaker Pipelines and its integration with SageMaker S tudio to manage the entire ML workflows. Use SageMaker ML Lineage T racking for the running history of experiments and for auditing and compliance verifications.

D. Use SageMaker Pipelines and its integration with SageMaker Experiments to manage the entire ML workflows. Use SageMaker Experiments for the running history of experiments and for auditing and compliance verifications.

✅ **Correct answer: D. Use SageMaker Pipelines and its integration with SageMaker Experiments to manage the entire ML workflows. Use SageMaker Experiments for the running history of experiments and for auditing and compliance verifications.**

SageMaker Pipelines provides native DAG-based workflow orchestration and visualization in Studio. SageMaker Experiments tracks trials, parameters, metrics, and artifacts for experiment history and governance. Together they offer

the fine-grained control and visibility requested. This pairing is the intended, fully managed solution for MLOps workflows on SageMaker.

Incorrect answers:

❌ **A. CodePipeline + Lineage Tracking** – CodePipeline is not optimized for ML DAGs or native SageMaker steps, and ML Lineage tracks artifact provenance rather than the full experiment run history and comparisons. This option misses the requested capabilities. It increases integration effort without first-class Pipelines features. It's not the best fit.

❌ **B. CodePipeline + Experiments** – While Experiments is correct for tracking runs, CodePipeline lacks the native ML DAG constructs and visualization that Pipelines provides. You'd have to glue custom actions and lose ML-specific UX. Operational overhead would be higher and features poorer.

❌ **C. Pipelines + Lineage Tracking** – Pipelines is correct for DAGs, but ML Lineage does not replace Experiments for run tracking and governance. You would lose rich experiment comparisons and trial management. This only partially satisfies the tracking requirement.

Question 17

A company needs to give its ML engineers appropriate access to training data. The ML engineers must access training data from only their own business group. The ML engineers must not be allowed to access training data from other business groups.

The company uses a single AWS account and stores all the training data in Amazon S3 buckets. All ML model training occurs in Amazon SageMaker.

Which solution will provide the ML engineers with the appropriate access?

A. Enable S3 bucket versioning.

B. Configure S3 Object Lock settings for each user.

C. Add cross-origin resource sharing (C ORS) policies to the S3 buckets.

D. Create IAM policies. Attach the policies to IAM users or IAM roles.

✅ **Correct answer: D. Create IAM policies. Attach the policies to IAM users or IAM roles.**

Fine-grained, per-group access to S3 data is enforced with IAM policies (and optionally S3 bucket/prefix conditions). You can scope permissions by path prefixes that map to business groups and attach the policies to roles the ML engineers assume. This works cleanly with SageMaker, which uses roles to access training data. It centralizes governance in the account as required.

Incorrect answers:

❌ **A. Enable S3 bucket versioning** – Versioning protects against unintended overwrites and deletions but does not constrain who can access what. It's a data protection feature, not an authorization mechanism. It doesn't address group-based isolation. Access would still be unrestricted without policies.

❌ **B. Configure S3 Object Lock** – Object Lock provides WORM retention and legal hold, not identity-based authorization. It's unrelated to limiting cross-group visibility. Enabling it would complicate data management without solving the access control need. It's not designed for RBAC.

❌ **C. Add CORS policies** – CORS governs browser access from web origins and has nothing to do with IAM-level data scoping for engineers. It does not prevent one group's credentials from listing or reading another group's prefixes. It's orthogonal to the requirement. IAM remains the correct control plane.

Question 18

A company is running ML models on premises by using custom Python scripts and proprietary datasets. The company is using PyT orch. The model building requires unique domain knowledge. The company needs to move the models to AWS.

Which solution will meet these requirements with the LEAST effort?

A. Use SageMaker built-in algorithms to train the proprietary datasets.

B. Use SageMaker script mode and premade images for ML frameworks.

C. Build a container on AWS that includes custom packages and a choice of ML frameworks.

D. Purchase similar production models through AWS Marketplace.

✅ **Correct answer: B. Use SageMaker script mode and premade images for ML frameworks.**

Script mode lets you bring existing PyTorch training code with minimal changes while using AWS-maintained framework images. You don't need to author a custom container unless you have nonstandard system dependencies. This path preserves domain-specific code and accelerates migration. It is typically the least effort for moving custom PyTorch scripts to SageMaker.

Incorrect answers:

❌ **A. Use SageMaker built-in algorithms** – Built-in algorithms are great for common tasks but won't accommodate bespoke PyTorch training logic and domain-specific pipelines. Migrating to a built-in would require extensive rewrites. This contradicts the "least effort" constraint. It sacrifices custom behavior.

❌ **C. Build a custom container** – Custom containers provide maximum flexibility but at higher operational and maintenance cost. For standard PyTorch code, AWS's premade images already cover most dependencies. Building and hardening your own image is unnecessary work here. Reserve this for truly atypical stacks.

❌ **D. Purchase similar models in Marketplace** – Off-the-shelf models won't embed your proprietary dataset nuances or domain tricks. Replacing internal models changes behavior and may not meet accuracy or compliance needs. It also doesn't migrate your codebase. This option doesn't solve the stated problem.

Question 19

An ML engineer receives datasets that contain missing values, duplicates, and extreme outliers. The ML engineer must consolidate these datasets into a single data frame and must prepare the data for ML.

Which solution will meet these requirements?

A. Use Amazon SageMaker Data Wrangler to import the datasets and to consolidate them into a single data frame. Use the cleansing and enrichment functionalities to prepare the data.

B. Use Amazon SageMaker Ground Truth to import the datasets and to consolidate them into a single data frame. Use the human-in-the-loop capability to prepare the data.

C. Manually import and merge the datasets. Consolidate the datasets into a single data frame. Use Amazon Q Developer to generate code snippets that will prepare the data.

D. Manually import and merge the datasets. Consolidate the datasets into the data.

✅ **Correct answer: A. Use Amazon SageMaker Data Wrangler to import the datasets and to consolidate them into a single data frame. Use the cleansing and enrichment functionalities to prepare the data.**

Data Wrangler provides a no/low-code UI with transforms for de-duplication, missing value handling, and outlier treatment. It unifies multiple sources into a single data flow and exports to Pandas/Spark pipelines for production. This minimizes custom coding while ensuring repeatability. It directly addresses consolidation and preparation in one place.

Incorrect answers:

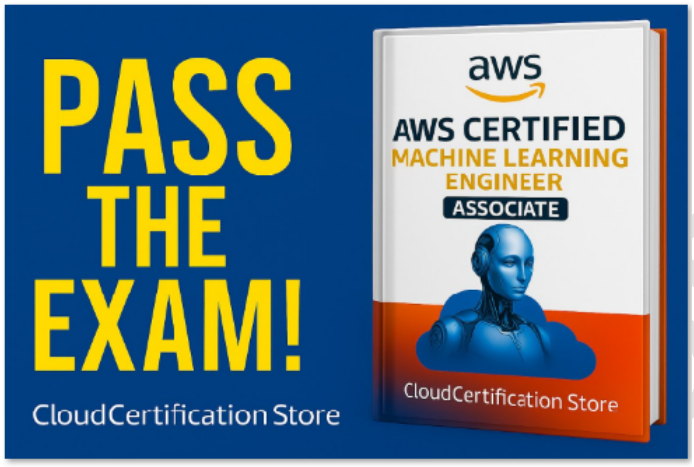
❌ **B. Use SageMaker Ground Truth** – Ground Truth is for data labeling, not for data cleaning, joining, and feature engineering. Human-in-the-loop workflows do

not address deduplication or outlier handling at scale. It's the wrong tool for this job. It increases cost and complexity without benefit.

❌ **C. Manual import + Amazon Q Developer snippets** – While feasible, this relies on ad-hoc code generation and manual stitching. It lacks the integrated preview, profiling, and export capabilities of Data Wrangler. Maintaining the code becomes the team's burden. It's higher effort and less robust.

❌ **D. Manually import and merge; incomplete option** – Even if completed, a purely manual pathway is more error-prone and harder to standardize. It provides no visual lineage of transforms. It doesn't leverage managed transforms or exportable pipelines. It fails the efficiency goal.

(END OF QUESTIONS)



PREVIEW COPY - PLEASE SHARE

Find the full 400+ questions document at

the Cloud Certification Store >> Click Here

All single exams are under $10. You can download free Previews as well.

As a value added bonus, use our courtesy discount code 25-NAY6XE6EEG

Final Review Checklist & Exam Readiness Scorecard

✅ How to Use the Final Review Checklist

This section is meant to **validate your hands-on skills and theoretical readiness** across all exam topics.

Step-by-step:

1. **Print it or load it in a note-taking app** (Notion, Google Docs, OneNote, etc.).
2. Go through each checkbox:
    - ✅ Check it if you **fully understand and can implement** the topic without looking up documentation.
    - ❌ Leave it unchecked if you feel unsure or haven't practiced the task.
3. Prioritize unchecked topics by reviewing:
    - Check the official documentation
    - Practice exams
    - Hands-on labs
4. For each **unchecked item**, write a short action plan or resource link next to it.

📈 How to Use the Exam Readiness Scorecard

This part helps you **self-assess your confidence level** and **focus your revision time** wisely.

Instructions:

1. For each domain (e.g., "Hybrid connectivity and routing"), **rate yourself** from 1 to 5:

   - 1️⃣ = No understanding or hands-on practice
   - 3️⃣ = Moderate familiarity, but need review
   - 5️⃣ = Mastered topic and can apply it in real-world use

2. Add **Notes / Action Items** to explain:
   - Why you scored yourself low
   - What resources you'll use to improve (YouTube, whitepapers, exam guides)
   - Practice test scores if relevant
3. Reassess **2–3 days before your exam**, and compare scores to measure improvement.

🧠 Bonus Tips

- Do **timed mock exams** and cross-reference errors with checklist topics
- Use the scorecard to **simulate an exam debrief**: where did you fail? What must you strengthen?

Once all checklist items are ✅ and all categories are at **4–5 stars** and you're consistently scoring **85%+** on full practice exams with confidence in scenario-based reasoning, then 🎯 you're likely **ready to book the real exam.**

Final Review Checklist

🧠 ML Concepts & Data Preparation

☐ Understand supervised, unsupervised, and reinforcement learning use cases
☐ Identify features, labels, training/test data, and sources of bias
☐ Apply data preprocessing: normalization, missing values, feature engineering
☐ Select correct algorithm for classification, regression, clustering, or recommendation

📦 Data Engineering on AWS

☐ Use S3 as a data lake with lifecycle policies and partitioning
☐ Build data ingestion pipelines with Kinesis, Glue, and Data Pipeline
☐ Use Athena, Glue Data Catalog, and Redshift Spectrum for queries
☐ Apply ETL/ELT transformations in Glue or EMR

⚙️ Model Training & Deployment

☐ Train models using Amazon SageMaker built-in algorithms and custom scripts
☐ Optimize models with hyperparameter tuning jobs and automatic model tuning
☐ Use SageMaker training jobs with managed Spot Training for cost efficiency
☐ Deploy models with SageMaker Endpoints (real-time, batch transform, multi-model)

🔍 Evaluation & Optimization

☐ Select evaluation metrics: accuracy, precision, recall, F1, RMSE, AUC
☐ Detect and address overfitting/underfitting (regularization, dropout, data augmentation)
☐ Apply cross-validation and holdout sets correctly
☐ Optimize inference latency and cost using instance types and autoscaling

🔐 Security & Compliance

☐ Configure IAM roles for SageMaker notebooks, training, and endpoints
☐ Protect data with KMS encryption (at rest & in transit)
☐ Manage secrets with Secrets Manager or Parameter Store
☐ Audit activity with CloudTrail and CloudWatch Logs

📊 Monitoring & Automation

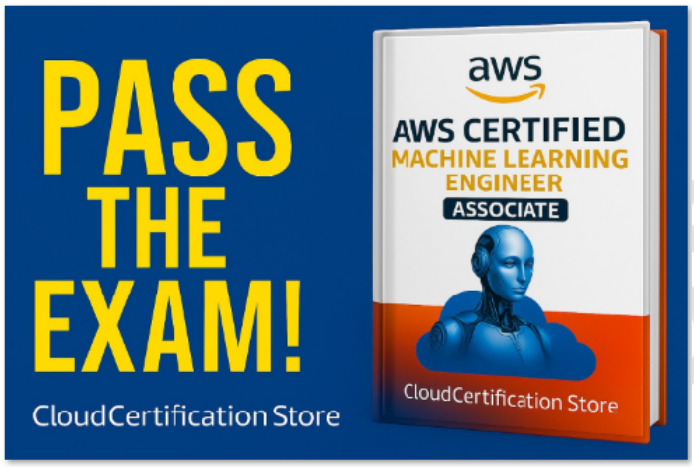
☐ Monitor training and inference jobs with CloudWatch metrics and logs
☐ Enable SageMaker Model Monitor for data drift and bias detection
☐ Automate retraining workflows with Step Functions and EventBridge
☐ Track experiments and lineage with SageMaker Experiments

Exam Readiness Scorecard

Domain	Confidence (1–5)	Notes / Action Items
🧠 ML concepts & data prep	☐1 ☐2 ☐3 ☐4 ☐5	
📦 Data engineering on AWS	☐1 ☐2 ☐3 ☐4 ☐5	
⚙️ Model training & deployment	☐1 ☐2 ☐3 ☐4 ☐5	
🔍 Model evaluation & optimization	☐1 ☐2 ☐3 ☐4 ☐5	
🔐 Security & compliance	☐1 ☐2 ☐3 ☐4 ☐5	
📊 Monitoring, automation & operations	☐1 ☐2 ☐3 ☐4 ☐5	
⏱️ Time management (130-min pacing)	☐1 ☐2 ☐3 ☐4 ☐5	Timed 65-question drill

💫 **Congratulations!! You are on the right path to certification.**

All of our practice exams include **300 + questions**. **This one in particular, contains way over 400 questions.**



PREVIEW COPY - PLEASE SHARE

Find the full 400+ questions document at

the Cloud Certification Store >> Click Here

All single exams are under $10. You can download free Previews as well.

As a value added bonus, use our courtesy discount code 25-NAY6XE6EEG

Our writers who have taken the exam recently—and the reviewers who purchased these materials—agree that **over 90 %** of the questions matched what they saw on the live test.

Invest in your future: browse the full catalogue of Cloud practice exams at our store

Featured Collection

2025 Google Cloud Professional Cloud Architect Exam Questions (GOOG-PCA-0010)

~~$8.89~~ $6.22

2025 Google Cloud Professional Cloud Developer Certification Practice Exam Questions (GOOG-PCD-0010)

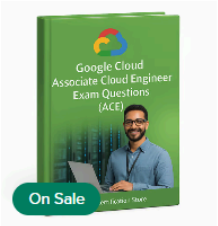
~~$8.89~~ $6.22

2025 Google Cloud Professional Machine Learning Engineer Practice Exam Questions (GOOG-PMLE-0010)

~~$8.99~~ $6.29

2025 Google Cloud Professional Data Engineer Certification Practice Exam Questions (GOOG-PDE-0010)

~~$8.89~~ $6.22

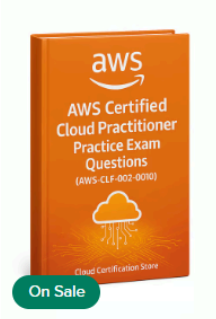


2025 Google Cloud Associate Cloud Engineer Exam Questions (GOOG-ACE-0010)

~~$8.89~~ $6.22

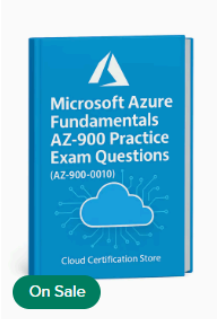
2025 Google Cloud Professional Cloud Database Engineer Practice Exam Questions (GOOG-PCDE-0010)

~~$8.89~~ $6.22



2025 AWS Certified Cloud Practitioner Practice Exam Questions (AWS-CLF-002-0010)

~~$8.99~~ $6.29



2025 Microsoft Azure Fundamentals AZ-900 Practice Exam Questions (AZ-900-0010)

~~$8.89~~ $6.22

2025 Microsoft Azure Developer AZ-204 Practice Exam Questions (AZ-204-0010)

~~$8.89~~ $6.22

FREE Practice Questions for the Google Cloud Generative AI Leader Exam

$0.00+

Google Professional

Machine Learning Engineer Practice Exam (PMLE)

(GOOG-PMLE-0010)

Cloud Certification Store

Google Professional Machine Learning Engineer

Cloud Certification Store

Google Cloud Professional Machine Learning Engineer - Practice Exam Questions (GOOG-PMLE-0010)

DISCLAIMER

- This practice test includes questions **compiled from various exam preparation platforms.**
- Important: **Questions and answers were AI-assisted and human-curated**. Verify accuracy with official Google documentation before relying on this material.
- **Users are strongly encouraged to** double-check all content against **official documentation and trusted sources** before using it for exam preparation or making important decisions.
- The creators of this material assume **no responsibility** for any errors, inaccuracies, or outcomes, including exam results, based on the use of this content.
- **Some questions might be duplicated or close** to previous ones, this is done on purpose as a way to re-inforce your learning.
- Single-user licence only
    - Includes one unique Payhip Licence Key per purchase, along with a Product Key.
    - Redistribution, resale, or public posting is prohibited. We can trace any file to the purchaser, with the use of the purchased License Key and Product Key.

Google Cloud Professional Machine Learning Enginner - Practice Questions (GOOG-PMLE-0010) - *PREVIEW 20 out of over 300 questions*

A Professional Machine Learning Engineer builds, evaluates, productionizes, and optimizes AI solutions by using Google Cloud capabilities and knowledge of conventional ML approaches. The ML Engineer handles large, complex datasets and creates repeatable, reusable code. The ML Engineer designs and operationalizes generative AI solutions based on foundational models. The ML Engineer considers responsible AI practices, and collaborates closely with other job roles to ensure the long-term success of AI-based applications. The ML Engineer has strong programming skills and experience with data platforms and distributed data processing tools. The ML Engineer is proficient in the areas of model architecture, data and ML pipeline creation, generative AI, and metrics interpretation. The ML Engineer is familiar with foundational concepts of MLOps, application development, infrastructure management, data engineering, and data governance. The ML Engineer enables teams across the organization to use AI solutions. By training, retraining, deploying, scheduling, monitoring, and improving models, the ML Engineer designs and creates scalable, performant solutions.

*Note: The exam does not directly assess coding skill. If you have a minimum proficiency in Python and Cloud SQL, you should be able to interpret any questions with code snippets.

The Professional Machine Learning Engineer exam assesses your ability to:

- ✔ Architect low-code AI solutions
- ✔ Collaborate within and across teams to manage data and models
- ✔ Scale prototypes into ML models
- ✔ Serve and scale models
- ✔ Automate and orchestrate ML pipelines
- ✔ Monitor AI solutions

This version of the Professional Machine Learning Engineer exam covers tasks related to generative AI, including building AI solutions using Model Garden and Vertex AI Agent Builder, and evaluating generative AI solutions.

To learn more about Google Cloud's generative AI services, go to Google Cloud Skills Boost to see the Introduction to Generative AI Learning Path (all audiences) or the Generative AI for Developers Learning Path (technical audience). If you are a partner, refer to the Gen AI partner courses: Introduction to Generative AI Learning Path , Generative AI for ML Engineers , and Generative AI for Developers. For additional learning, refer to product-specific Gen AI learning offerings, such as Explore and Evaluate Models using Model Garden , Vertex AI Agent Builder path (partners), and Integrate Search in Applications using Vertex AI Agent Builder .

About this certification exam

Length: Two hours

Registration fee: $200 (plus tax where applicable)

Language: English

Exam format: 50-60 multiple choice and multiple select questions

Exam delivery method:

a. Take the online-proctored exam from a remote location, review the online testing requirements.

b. Take the onsite-proctored exam at a testing center, locate a test center near you

Prerequisites: None

Recommended experience: 3+ years of industry experience including 1 or more years designing and managing solutions using Google Cloud.

Certification Renewal / Recertification: Candidates must recertify in order to maintain their certification status. Unless explicitly stated in the detailed exam descriptions, all Google Cloud certifications are valid for two years from the date of certification. Recertification is accomplished by retaking the exam during the recertification eligibility time period and achieving a passing score. You may attempt recertification starting 60 days prior to your certification expiration date.

Exam overview

Step 1: Get real world experience

Before attempting the Machine Learning Engineer exam, it's recommended that you have 3+ years of hands-on experience with Google Cloud products and solutions. Ready to start building? Explore the Google Cloud Free Tier for free usage (up to monthly limits) of select products.

Try the Google Cloud Free Tier

Step 2: Understand what's on the exam

The exam guide contains a complete list of topics that may be included on the exam. Review the exam guide to determine if your skills align with the topics on the exam.

See current exam guide

Step 3: Review the sample questions

Familiarize yourself with the format of questions and example content that may be covered on the Machine Learning Engineer exam.

Review sample questions

Step 4: Round out your skills with training

Collapse all

Follow the learning path

Prepare for the exam by following the Machine Learning Engineer learning path. Explore online training, in-person classes, hands-on labs, and other resources from Google Cloud.

Start preparing

Take a webinar

Prepare for the exam with Googlers and certified experts. Get valuable exam tips and tricks, as well as insights from industry experts.

Sign up

Additional resources

Explore Google Cloud documentation for in-depth discussions on the concepts and critical components of Google Cloud.

Learn about designing, training, building, deploying, and operationalizing secure ML applications on Google Cloud using the Official Google Cloud Certified Professional

Machine Learning Engineer Study Guide. This guide uses real-world scenarios to demonstrate how to use the Vertex AI platform and technologies such as TensorFlow, Kubeflow, and AutoML, as well as best practices on when to choose a pretrained or a custom model.

Step 5: Schedule an exam

Register and select the option to take the exam remotely or at a nearby testing center.

Review exam terms and conditions and data sharing policies.

Take the next step

Follow the learning path: Start Learning

Earn a skill badge in machine learning
Start now

New to Google Cloud?
Get started

Take a cert prep webinar
Watch Cloud OnAir

Practice Questions

Question 1

You built a custom ML model using scikit-learn. Training time is taking longer than expected. You decide to migrate your model to Vertex AI Training, and you want to improve the model's training time. What should you try out first?

A. Train your model in a distributed mode using multiple Compute Engine VMs.
B. Train your model using Vertex AI Training with CPUs.
C. Migrate your model to TensorFlow, and train it using Vertex AI Training.
D. Train your model using Vertex AI Training with GPUs.

✅ **Correct answer: B. Train your model using Vertex AI Training with CPUs**
📌 **Explanation:** Before making major changes to your model or infrastructure, start by running your scikit-learn training on Vertex AI's optimized environment using standard CPU instances. Vertex AI's deep learning VM images come with optimized math libraries (like Intel MKL) for NumPy/SciPy, which can accelerate scikit-learn computations.

This approach requires minimal code changes and leverages efficient vectorized operations, potentially speeding up training significantly. It's a simpler first step than rewriting your model in TensorFlow or implementing complex distributed training, and it lets you gauge performance improvements from a tuned environment. If training is still slow, you can then consider more involved options (like GPUs or distributed training).

Incorrect answers:

❌ **A. Train in distributed mode on multiple VMs** – Running scikit-learn algorithms across multiple machines is non-trivial. Scikit-learn doesn't natively distribute training across nodes, so this would add complexity and overhead with uncertain benefit. It's not the first thing to try when a simpler environment change (optimized single-node training) might suffice.

❌ **C. Migrate to TensorFlow and retrain** – Converting your scikit-learn model to TensorFlow would be time-consuming and isn't guaranteed to solve the speed issue. You'd only consider

this if you needed GPU acceleration or distributed training that scikit-learn can't handle. It's a heavy lift to rewrite the model, so it shouldn't be the first step.

❌ **D. Use GPUs on Vertex AI Training** – Many scikit-learn algorithms do not automatically benefit from GPU acceleration, as they run on CPU by design. Without specialized libraries, simply switching to a GPU instance might not improve training time for scikit-learn models. It's better to first optimize on CPUs. (If your algorithm can utilize GPU-enabled libraries, you might explore this later, but it's not the default assumption for scikit-learn.)

Question 2

You work for a gaming company that has millions of customers worldwide. All games offer a chat feature allowing real-time communication in over 20 languages, translated on-the-fly via the Cloud Translation API. You built an ML system to automatically moderate chat messages for toxicity. However, the model's performance varies greatly by language – it fails more often on certain languages. The factory has no reliable internet, and faster defect detection is a priority. (Your company wants to implement the new ML model ASAP.) How should you improve the model's performance across languages?

A. Add a regularization term (e.g., MinDiff) to the loss function to reduce bias.
B. Train a separate classifier using the chat messages in their original languages (no translation).
C. Replace the in-house word2vec embeddings with a large multilingual model like GPT-3 or T5.
D. Remove moderation for languages where the false positive rate is too high.

✅ **Correct answer: D. Remove moderation for languages for which the false positive rate is too high**

📌 **Explanation:** The simplest way to ensure uniform user experience when certain languages are yielding many moderation errors is to disable automated moderation for those languages. If the model is over-flagging benign content in underrepresented languages, turning off or limiting the model's actions on those languages prevents unfair blocking of users. This approach immediately eliminates the model's false positives for those languages, achieving uniform (if

conservative) treatment across all languages without requiring infrastructure changes or long development cycles. Given the urgent need and lack of time to retrain a better model, focusing moderation on languages where the model is reliable and omitting it where it's unreliable is a practical stop-gap solution.

Incorrect answers:

❌ **A. Add MinDiff regularization** – MinDiff is used to reduce *unfair bias* by penalizing performance gaps between predefined groups, but it requires defining those groups and retraining the modelruslanmv.com. This is a complex change and wouldn't quickly fix the immediate issue of poor performance on certain languages, especially if training data for those languages is limited.

❌ **B. Train on original languages** – Building separate models or a single multilingual model that processes messages in their original language could improve accuracy, but it's time and resource intensive. It requires either collecting labeled data for each language or using multilingual NLP techniques. The question scenario emphasizes implementing a solution *as soon as possible*, so retraining a new multi-language model is not the fastest remedy.

❌ **C. Use GPT-3 or T5 embeddings** – Swapping in a massive pretrained model (like GPT-3/T5) for embeddings could improve cross-language understanding, but those models are extremely resource-heavy. Using them would likely violate the "no infrastructure change" constraint and add significant latency or cost. Additionally, integrating such models is a non-trivial engineering effort. This option is neither quick nor cost-effective for an immediate fix.

Question 3

You need to ensure real-time ingestion of user activity data from a mobile app into BigQuery for analysis and ML experimentation. Your team will use BigQuery for data analysis and ML modeling. What should you do to ingest the streaming data into BigQuery with minimal latency?

A. Configure Pub/Sub to stream the data into BigQuery.
B. Run an Apache Spark streaming job on Dataproc to ingest the data into BigQuery.

C. Run a Dataflow streaming job to ingest the data into BigQuery.

D. Use Pub/Sub together with a Dataflow streaming job to ingest the data into BigQuery.

✅ **Correct answer: A. Configure Pub/Sub to stream the data into BigQuery**

📌 **Explanation:** The easiest and most efficient solution is to use **Pub/Sub's BigQuery subscription feature**, which writes messages directly from Pub/Sub into a BigQuery table. This native integration enables real-time streaming inserts with minimal setupinfoq.cominfoq.com. It eliminates the need to manage a separate processing job, providing low-latency ingestion out-of-the-box. By configuring a Pub/Sub topic for the app events and creating a BigQuery subscription on that topic, the data will flow continuously into BigQuery as it arrives, satisfying the real-time requirement.

Incorrect answers:

❌ **B. Use Spark on Dataproc** – Setting up a Spark streaming job introduces unnecessary complexity. You'd have to manage a Spark cluster and code the ingestion logic. This is heavyweight for simply piping events into BigQuery, and it likely adds more latency compared to Google's native streaming mechanisms.

❌ **C. Use Dataflow streaming** – While Dataflow can stream data to BigQuery, it's an extra layer here. If no transformation is needed on the data, using Dataflow is overkill for just shuttling data from Pub/Sub to BigQuery. The Pub/Sub→BigQuery direct path is simpler and managed for you, whereas Dataflow would require developing and maintaining a pipeline (and D is an even more explicit version of this approach).

❌ **D. Pub/Sub with Dataflow** – This classic combination (Pub/Sub as source, Dataflow pipeline to sink into BigQuery) is traditionally used for streaming ingestion, but Google's newer **BigQuery subscription** removes the need for the Dataflow step. Option D would work, but it's not the quickest or most efficient route. It incurs more cost and maintenance effort compared to simply letting Pub/Sub forward messages to BigQuery directly.

Question 4

You recently trained a deep learning model using Keras on a large dataset. After a few epochs, you notice the training and validation losses barely change – the model isn't learning. You want to debug and fix your model quickly. What should you do first?

A. Verify that your model can achieve a low loss on a small subset of the dataset.
B. Add handcrafted features to inject your domain knowledge into the model.
C. Use Vertex AI's hyperparameter tuning service to find a better learning rate.
D. Switch to hardware accelerators and train the model for more epochs.

✅ **Correct answer: A. Verify that your model can obtain a low loss on a small subset of the dataset**

📌 **Explanation:** When a model is not learning (loss not decreasing), a key first step is to **check for fundamental issues by training on a very small sample**. If your model can't overfit a tiny dataset, something is likely wrong with the model architecture, data processing, or training setup. By using a small subset (even just a few batches) and seeing if the model can drive loss near zero, you verify that the training pipeline is working at a basic level. This fast sanity check helps distinguish between a model capacity/optimization problem and a bug or data issue. It's a quick, low-effort diagnostic step to perform before trying more complex fixes.

Incorrect answers:

❌ **B. Add handcrafted features** – Introducing manual features is not the first line of action for a non-learning model. If the model isn't learning from existing features at all (loss plateaus immediately), adding features won't help until you resolve why learning stalled (which could be due to bugs, normalization issues, etc.). Handcrafted features are more relevant if the model is learning but not achieving required accuracy, not when it's basically stuck.

❌ **C. Hyperparameter tuning for learning rate** – While an improper learning rate can cause training issues (too high can cause oscillation, too low can stall learning), if losses are *barely changing at all*, it might indicate a more fundamental issue (like a bug). It's wise to first verify the model can learn in a simple scenario. Only after that would you systematically tune hyperparameters. Jumping straight to automated tuning without basic debugging could waste time.

❌ **D. Use accelerators and train longer** – If the model isn't improving, simply training for more epochs or on faster hardware won't magically fix it. Faster hardware (GPUs/TPUs) just does the same computations quicker; it doesn't address why loss is stagnant. Without diagnosing the root cause, you might just more quickly reach the same plateau. It's better to troubleshoot with short, controlled experiments (like option A) before scaling up compute.

Question 5

You are experimenting with an XGBoost classification model in Vertex AI Workbench. You split your BigQuery data into training and validation sets using these SQL queries:

pgsql

```
CREATE OR REPLACE TABLE training AS
(SELECT * FROM mytable WHERE RAND() <= 0.8);

CREATE OR REPLACE TABLE validation AS
(SELECT * FROM mytable WHERE RAND() <= 0.2);
```

After training, your model's AUC ROC is 0.80 on validation, but when deployed in production its AUC drops to 0.65. What is the most likely cause?

A. There is training-serving skew in your production environment.
B. The training dataset was too small to generalize well.
C. The training and validation tables share some records, meaning not all data was used properly.
D. The RAND() function caused every record in validation to also be in training.

✅ **Correct answer: A. There is training-serving skew in your production environment.**
📌 **Explanation:** The sudden performance drop from validation to production strongly suggests **training-serving skew**, meaning the model is seeing data in production that has a different

distribution or features than the data it was validated onfreecram.net. In this scenario, the method of splitting data is flawed: using separate `RAND()` filters can lead to overlapping datasets (many records appear in both training and validation)freecram.netfreecram.net. This overlap would make validation metrics overly optimistic (since the model inadvertently "saw" validation data during training), thus the high 0.80 AUC. Once deployed, the model faces truly unseen data and its performance (0.65 AUC) reflects the true generalization ability. This discrepancy – excellent metrics during validation vs poor in production – is a classic symptom of training-serving skew or data leakage.

Incorrect answers:

❌ **B. Training dataset insufficient** – If data volume were the issue, you'd likely see high variance (unstable metrics) or overfitting on training vs validation. Here the validation looked good, but production did not, indicating a shift in data rather than simply not enough data. Also, nothing in the scenario suggests an extremely small dataset; the issue is how the split was done.

❌ **C. Training and validation tables share some records** – This is true (they *do* share records given the RAND logic), but the option says "not using all data in initial table," which is a bit off. The bigger problem is the overlap itself causing misleading validation results, which is essentially part of training-serving skew. However, answer A names the broader issue more accurately. (Option C is on the right track but doesn't explicitly connect to the observed performance drop in production; the key issue is the skewed evaluation.)

❌ **D. Every validation record is in training** – It's an exaggeration to say *every* record overlaps (though theoretically any record with RAND() ≤ 0.2 would also satisfy ≤ 0.8). In practice, a large portion (roughly 20% of data) ended up in both setsfreecram.net. This data leakage is problematic, but stating it as "RAND caused every record to be in both" is not strictly accurate. The root cause is the splitting method leading to overlapping data, which falls under training-serving skew. Thus A is the more encompassing description of the problem.

Question 6

During batch training of a neural network, you notice the loss oscillates (fluctuates up and down) instead of steadily decreasing. What adjustment will most likely ensure the model converges?

A. Decrease the size of each training batch.
B. Decrease the learning rate hyperparameter.
C. Increase the learning rate hyperparameter.
D. Increase the size of the training batch.

✅ **Correct answer: B. Decrease the learning rate hyperparameter.**
📌 **Explanation:** Loss oscillation during training is a strong sign that the learning rate is too high, causing the optimizer to overshoot minima and bounce around the loss surfacefreecram.netfreecram.net. By lowering the learning rate, you make the weight updates smaller and more fine-grained, which helps the model descend the loss curve more smoothly rather than jumping back and forthfreecram.net. A smaller learning rate often stabilizes training and allows the network to converge to a minimum. This is a common remedy for oscillating or diverging losses in neural network trainingmattermodeling.stackexchange.com.

Incorrect answers:

❌ **A. Decrease batch size** – Reducing batch size increases the noisiness of the gradient estimates, which generally causes *more* oscillation in the loss, not less. A very small batch might even worsen convergence or require an even smaller learning rate. The primary cause of oscillation here is likely the step size (learning rate), not the batch size.

❌ **C. Increase learning rate** – This would exacerbate the problem. If the loss is already oscillating, a higher learning rate would cause even larger weight updates, potentially making the training diverge completely (loss blowing up). It's the opposite of the needed adjustment.

❌ **D. Increase batch size** – While larger batches give more stable gradients (reducing noise), simply using a bigger batch does not directly fix oscillation caused by an overly large learning rate. There's usually an optimal batch size for throughput, but convergence issues are more effectively addressed by tuning learning rate or using techniques like momentum/adaptive optimizers. In fact, if oscillation is severe, you'd still need to lower the learning rate even with a bigger batch.

Question 7

You are working on a predictive maintenance model for factory machines. It's a binary classifier that predicts whether a crucial machine will fail in the next 3 days (class "1" means failure is predicted). Only 4% of training examples are actual failures, so missing a failure is far more costly than a false alarm. You evaluate several models on a test set and want to choose the one that **prioritizes detection** of failures while ensuring that more than 50% of its failure predictions are correct. Which model performance criterion should you prioritize in selecting the model?

A. The model with the highest AUC ROC, given its precision is above 0.5.
B. The model with the lowest RMSE and recall above 0.5.
C. The model with the highest recall, with precision above 0.5.
D. The model with the highest precision, with recall above 0.5.

✅ **Correct answer: C. The model with the highest recall, with precision above 0.5.**
📌 **Explanation: Recall** (sensitivity) measures how many of the actual failures your model catches. "Prioritizing detection" means you want to catch as close to 100% of failures as possible, so high recall is the top priority. However, you also require that when the model does predict a failure, it's correct more than half the time – in other words, precision > 50%. Among models that meet the precision > 0.5 threshold, you should pick the one with the highest recall. This ensures you're maximizing the captured failures (fewer missed failures), while keeping false alarms to a manageable rate (at least half of alarms are true issues).

Incorrect answers:

❌ **A. Highest AUC with precision > 0.5** – AUC ROC is a useful overall metric, but it doesn't directly address the operating threshold or the trade-off between missing failures and raising false alarms. A model could have a great AUC yet still have subpar recall at the chosen threshold. The specific requirement here is about recall and precision, not the entire ROC curve. Focusing on AUC could lead to choosing a model that performs well on average but not optimally for catching failures.

❌ **B. Lowest RMSE with recall > 0.5** – RMSE (Root Mean Squared Error) is a regression metric and not applicable to classification performance. Including RMSE doesn't make sense in this context. We need classification metrics like precision/recall, not a regression error metric.

❌ **D. Highest precision with recall > 0.5** – This would choose a model that makes very few false alarms (high precision), as long as it at least catches half of failures (recall > 0.5). But the problem stated that missing a failure is very costly, implying recall is more critical. A model that is extremely precise might be too conservative and miss many failures (as long as it catches just over 50%, it passes the recall > 0.5 check). That's not desirable here – we'd rather tolerate more false alarms if it means catching nearly all failures. Thus, recall should be the primary focus.

Question 8

You have a highly imbalanced dataset: 96% of the images do **not** contain your company's logo, and only 4% do. You're training a binary image classifier to detect the presence of the logo. Which evaluation metric will give you the most confidence that the model is performing well?

A. Precision
B. Recall
C. RMSE
D. F₁ score

✅ **Correct answer: D. F₁ score**
📌 **Explanation:** The **F₁ score** is the harmonic mean of precision and recall, and it's well-suited for imbalanced classification problems where you care about both false positives and false negatives. In this scenario, if you rely on accuracy, a trivial model that always predicts "no logo" would be 96% accurate but useless. The F₁ score will only be high if the model achieves a good balance of precision and recall – meaning it finds a reasonable fraction of logos *and* doesn't flood you with false alarms. This gives a more informative picture of performance on the minority class than either precision or recall alone. Essentially, a strong F₁ indicates the model is handling the skewed dataset well by capturing logos without too many mistakes.

Incorrect answers:

❌ **A. Precision** – Precision alone tells you, "When the model says *logo present*, how often is it correct?" A high precision could be achieved by a model that rarely ever flags a logo (it might miss most logos but be right on the few it detects). In an imbalanced setting, focusing solely on precision might lead to a model that ignores many positive cases. It doesn't ensure logos are being detected adequately.

❌ **B. Recall** – Recall tells you "What percentage of actual logos did the model detect?" High recall alone could be achieved by flagging lots of images (catching most logos but also producing many false positives). By itself, recall doesn't account for precision. In practice, you want both: find the logos (recall) and be accurate when you flag one (precision). That's why F₁ is preferred – it combines both aspects.

❌ **C. RMSE** – Root Mean Squared Error is a regression metric, not used for evaluating classification performance. It doesn't apply here since the task is categorical (logo vs no logo). Even if one treats the problem as predicting probabilities, other metrics like log-loss or AUC would be more relevant than RMSE. In short, RMSE is the wrong tool for classification.

Question 9

You are creating an ML pipeline to ingest hundreds of millions of rows from BigQuery and train TensorFlow models on Vertex AI. You want to minimize data ingestion bottlenecks and ensure the solution is scalable. What is the best way to feed the BigQuery data into your TensorFlow training job?

A. Use the BigQuery Python client to download the data into a pandas DataFrame, then use `tf.data.Dataset.from_tensor_slices()` on it.

B. Export the BigQuery data to CSV files in Cloud Storage, then use `tf.data.TextLineDataset()` to read the CSVs.

C. Convert the BigQuery data to TFRecord format, store in Cloud Storage, then use `tf.data.TFRecordDataset()` to read it.

D. Use **TensorFlow I/O's BigQuery Reader** to read directly from BigQuery into the training pipeline.

✅ **Correct answer: D. Use TensorFlow I/O's BigQuery Reader to directly read the data.**

📌 **Explanation: TensorFlow I/O's BigQuery integration** allows you to stream data from BigQuery to TensorFlow efficiently, without intermediate storage. This is the most scalable and maintenance-free option for large datasets. It avoids the overhead of exporting or downloading data and leverages BigQuery's ability to supply data in parallel to your training job. In short, option D provides a direct, end-to-end pipeline with minimal bottlenecks – it was designed for exactly this purpose, enabling TensorFlow to consume BigQuery data in a distributed manner.

Incorrect answers:

❌ **A. BigQuery client to DataFrame, then from_tensor_slices** – Loading hundreds of millions of rows into a pandas DataFrame will likely run out of memory or be painfully slow. Even if it succeeded, wrapping a DataFrame with `from_tensor_slices` would force all data into memory and not leverage streaming or parallel prefetching. This approach doesn't scale to very large data.

❌ **B. Export to CSV and use TextLineDataset** – This introduces an extra step (exporting) and large CSV files, which are bulky to parse. Reading CSVs line by line is relatively slow compared to using a binary format. It also incurs additional storage and I/O overhead. It's a viable workaround for smaller datasets, but not the most efficient for massive data.

❌ **C. Use TFRecords via Cloud Storage** – TFRecord is a binary format that is efficient for TensorFlow, and using it would be scalable. However, the process of converting "hundreds of millions" of rows to TFRecords still means you have to run an export job (or Dataflow pipeline) to produce those records. That's an extra heavy step to manage. Option D skips the conversion and directly feeds data to training, simplifying the workflow. If you already had data in TFRecords, great – but here the data lives in BigQuery, so reading it directly is faster and easier than building a separate TFRecord export pipeline.

Question 10

You manage server maintenance at a data center and want to build a predictive maintenance model to detect machines likely to fail. You have lots of sensor data, but **no labeled incidents** (no examples of "failure" vs "normal" labeled yet). What is the first thing you should do?

A. Train a time-series model to predict each machine's performance metrics and trigger an alert when actual readings deviate significantly from predictions.
B. Develop a simple rule (e.g. a z-score threshold) to label historical data as "anomalous" or not, and use that rule in real-time to monitor machines.
C. Use a simple rule (e.g. z-score) on historical data to create labels, then train a supervised model on that newly labeled dataset to detect anomalies.
D. Hire experts to manually label the historical sensor data for past failures, then train a model on those labels.

✅ **Correct answer: A. Train a time-series model to predict the machines' performance values, and alert on significant deviations.**
📌 **Explanation:** With no labeled failure data, the best starting point is an **unsupervised or self-supervised anomaly detection approach**. Training a time-series forecasting model on each machine's normal behavior, then flagging large prediction errors, is a common technique. Essentially, the model learns the expected patterns (temperature, vibration, etc.), and if a machine's actual readings diverge sharply from the model's prediction, it likely indicates a potential issue. This approach doesn't require explicit failure labels and can start providing actionable insights immediately. It prioritizes catching any unusual behavior (potential failures) without needing a labeled dataset upfront.

Incorrect answers:

❌ **B. Heuristic rule for real-time monitoring** – Using a simple statistical threshold (like "if metric > 3σ from mean, flag it") might catch gross anomalies, but it's a very crude method. It doesn't learn the nuanced patterns of each machine or adapt over time. Also, by itself it doesn't improve your understanding of the data or create a model – it's just a one-off rule. It's often better to incorporate such heuristics into a model or use them to validate a model's alerts, rather than as the primary solution.

❌ **C. Label with heuristic then train a model** – This is effectively two steps: use a rule to generate pseudo-labels for anomalies, then train a supervised model on those labels. The issue is that if your initial heuristic is poor, your model will just learn that flawed heuristic. It's a bit circular. Given no true labels, it's usually more effective to directly use unsupervised anomaly detection (like option A) than to assume your made-up labels are ground truth. That said, once you have some confidence (or some actual failure examples), you could refine with supervised learning – but that's not the *first* thing to do.

❌ **D. Manually label historical data** – In predictive maintenance, true failures might be rare, and it may not be even possible for humans to correctly label "when did this sensor reading indicate a future failure?" without hindsight. Moreover, it's time-consuming and costly to have analysts label tons of sensor logs, and they might still miss subtle precursors to failure. It's better initially to use algorithms to detect anomalies. Human labeling could come into play later, once you have alerts, to confirm which alerts corresponded to real issues and then improve the model. But at the outset, it's impractical to label everything.

Question 11

You are experimenting with a built-in distributed XGBoost model in Vertex AI Workbench user-managed notebooks. You use BigQuery to split your data into training and validation sets using the following queries:

```
CREATE OR REPLACE TABLE ‘myproject.mydataset.training‘ AS
 (SELECT * FROM ‘myproject.mydataset.mytable‘ WHERE RAND() <= 0.8);

CREATE OR REPLACE TABLE ‘myproject.mydataset.validation‘ AS
 (SELECT * FROM ‘myproject.mydataset.mytable‘ WHERE RAND() <= 0.2);
```

After training the model, you achieve an area under the receiver operating characteristic curve (AUC ROC) value of 0.8, but after deploying the model to production, you notice that your model performance has dropped to an AUC ROC value of 0.65. What problem is most likely occurring?

A. There is training-serving skew in your production environment.
B. There is not a sufficient amount of training data.
C. The tables that you created to hold your training and validation records share some records, and you may not be using all the data in your initial table.
D. The RAND() function generated a number that is less than 0.2 in both instances, so every record in the validation table will also be in the training table.

Correct answer: A. There is training-serving skew in your production environment.

📌 AUC ROC degradation from 0.8 to 0.65 in production typically indicates that the data seen at serve time differs from the training distribution—i.e., training-serving skew Google - Professional M….

Incorrect answers:

❌ B. There is not a sufficient amount of training data. – New data would impact both training and production, not just serving.
❌ C. The tables … share some records … – That causes validation leakage, which would inflate validation performance, not degrade production relative to training.
❌ D. The RAND() function … – While possible, this would affect both training and validation splits, not cause skew between training and serving.

Question 12

Your team needs to build a model that predicts whether images contain a driver's license, passport, or credit card. The data engineering team already built the pipeline and generated a dataset composed of 10,000 images with driver's licenses, 1,000 images with passports, and 1,000 images with credit cards. You now have to train a model with the following label map: [`drivers_license`, `passport`, `credit_card`]. Which loss function should you use?

A. Categorical hinge
B. Binary cross-entropy

C. Categorical cross-entropy
D. Sparse categorical cross-entropy

✅ **Correct answer: C. Categorical cross-entropy**
📌 For a multiclass classification problem with three mutually exclusive classes, use categorical cross-entropy. It measures the difference between the true one-hot label distribution and the predicted probability distribution over all classes.

Incorrect answers:
❌ **A. Categorical hinge** – Used for "max-margin" loss in SVM-style multiclass; not standard for neural networks.
❌ **B. Binary cross-entropy** – Applies to independent binary labels (multi-label), not mutually exclusive multiclass.
❌ **D. Sparse categorical cross-entropy** – Also multiclass, but expects integer class IDs rather than one-hot encoded vectors. The question implies one-hot encoding via a label map.

Question 13

You are an ML engineer at a manufacturing company. You need to build a model that identifies defects in products based on images taken at the end of the assembly line. You want your model to preprocess the images with lower computation to quickly extract features of defects. Which approach should you use to build the model?

A. Reinforcement learning
B. Recommender system
C. Recurrent Neural Networks (RNN)
D. Convolutional Neural Networks (CNN)

✅ **Correct answer: D. Convolutional Neural Networks (CNN)**
📌 CNNs are purpose-built for image processing. Their convolutional layers extract spatial features efficiently, enabling fast inference on defects.

Incorrect answers:

❌ **A. Reinforcement learning** – For sequential decision making, not static image classification.

❌ **B. Recommender system** – For suggesting items or content, not image analysis.

❌ **C. RNN** – Designed for sequential data (text/time series), not spatial feature extraction in images.

Question 14

You are developing an ML model intended to classify whether X-ray images indicate bone fracture risk. You have trained a ResNet architecture on Vertex AI using a TPU accelerator, but you're unsatisfied with training time and memory usage. You want to quickly iterate your training code without major changes and minimize impact on accuracy. What should you do?

A. Reduce the number of layers in the model architecture.
B. Reduce the global batch size from 1024 to 256.
C. Reduce the dimensions of the images used in the model.
D. Configure your model to use bfloat16 instead of float32.

✅ **Correct answer: D. Configure your model to use bfloat16 instead of float32**

📌 Switching to bfloat16 halves memory usage and doubles throughput on TPUs with minimal code change. Model accuracy remains nearly unchanged because bfloat16 preserves dynamic range.

Incorrect answers:

❌ **A. Reduce layers** – Alters model capacity and may degrade accuracy; requires code refactoring.

❌ **B. Reduce batch size** – Eases memory pressure but slows overall throughput and convergence dynamics.

❌ **C. Reduce image dimensions** – May lose critical diagnostic details, risking accuracy.

Question 15

You have successfully deployed to production a complex TensorFlow model trained on tabular data. You want to predict the lifetime value (LTV) field for each subscription stored in the BigQuery table subscription.subscriptionPurchase in project my-fortune500-company-project. You organized your entire TFX pipeline (preprocessing → validation → training → deployment) as a Vertex AI pipeline. You need to prevent prediction drift—i.e., feature distributions changing significantly over time—without retraining too often. What should you do?

A. Implement continuous retraining of the model daily using Vertex AI Pipelines.
B. Add a model monitoring job where 10% of incoming predictions are sampled once every 24 hours.
C. Add a model monitoring job where 90% of incoming predictions are sampled once every 24 hours.
D. Add a model monitoring job where 10% of incoming predictions are sampled every hour.

✅ **Correct answer: D. Add a model monitoring job where 10% of incoming predictions are sampled every hour.**
📌 Hourly sampling at 10% balances cost and timeliness, enabling you to detect feature-drift quickly so you can trigger retraining only when necessary.

Incorrect answers:
❌ **A. Continuous daily retraining** – High cost and unnecessary if data distributions haven't shifted.
❌ **B. 10% sampled once per day** – May delay detection of drift by up to 24 hours, risking stale predictions.
❌ **C. 90% sampled daily** – High cost for little extra benefit; sampling rate can remain low if frequency is high.

Question 16

You recently developed a deep learning model using Keras, experimenting with training strategies. You first trained on a single GPU, but it was too slow. Next you distributed training

across 4 GPUs using `tf.distribute.MirroredStrategy` without other changes, but saw no speedup. What should you do?

A. Distribute the dataset with `tf.distribute.Strategy.experimental_distribute_dataset`
B. Create a custom training loop.
C. Use a TPU with `tf.distribute.TPUStrategy`
D. Increase the batch size.

✅ **Correct answer: A. Distribute the dataset with `tf.distribute.Strategy.experimental_distribute_dataset`**
📌 MirroredStrategy requires that you explicitly shard and distribute your input dataset. Calling `experimental_distribute_dataset` ensures each GPU gets a slice of each batch, enabling true parallelism without code structure changes.

Incorrect answers:
❌ **B. Custom training loop** – Unnecessary complexity; the standard Keras loop works once data is distributed correctly.
❌ **C. TPU with TPUStrategy** – A different accelerator; doesn't address why the GPUs weren't utilized.
❌ **D. Increase batch size** – Might help GPU utilization, but if the data isn't distributed, batch size has no effect on multi-GPU scaling.

Question 17

You work for a gaming company that develops massively multiplayer online (MMO) games. You built a TensorFlow model that predicts whether players will make in-app purchases of > $10 in the next two weeks. The model's predictions adapt each user's game experience. User data is stored in BigQuery. How should you serve your model while optimizing cost, user experience, and ease of management?

A. Import the model into BigQuery ML. Make predictions using batch reads from BigQuery, then push results to Cloud SQL.
B. Deploy the model to Vertex AI Prediction. Make predictions using batch reads from Cloud Bigtable, then push results to Cloud SQL.
C. Embed the model in the mobile application. Make predictions after each in-app purchase event is published to Pub/Sub, then push results to Cloud SQL.
D. Embed the model in a streaming Dataflow pipeline. Make predictions after each in-app purchase event is published to Pub/Sub, then push results to Cloud SQL.

✅ **Correct answer: D. Embed the model in a streaming Dataflow pipeline.**
📌 A Dataflow streaming pipeline can consume Pub/Sub events in real time, call the TensorFlow model for low-latency predictions, and write enriched results to Cloud SQL. It keeps infrastructure serverless and managed, providing scale and minimal client-side complexity.

Incorrect answers:
❌ **A. BigQuery ML batch** – Imposes high latency; not real time.
❌ **B. Vertex AI batch on Bigtable** – Also batch, not suitable for real-time adaptation.
❌ **C. Embedding in mobile app** – Exposes model code to clients, complicates updates and security.

Question 18

You are building a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. The model uses a city name variable as a key predictive component. To train and serve the model, data must be in columns. You want the least coding while preserving predictive power. What should you do?

A. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file, and upload it as part of your model to BigQuery ML.
B. Create a new BigQuery view that omits city information.
C. Use Cloud Data Fusion to assign each city to a region labeled 1–5, then use that number as the city feature.
D. Use Dataprep to one-hot encode the state column and make each city a binary column.

✅ **Correct answer: A. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file, and upload it as part of your model to BigQuery ML.**

📌 BigQuery ML supports supplying a GCS-hosted vocabulary file for STRING features. You define a CSV of all city names, upload it, and BigQuery ML automatically one-hot encodes using that vocabulary—no ETL pipelines or manual column explosion required.

Incorrect answers:

❌ **B. Drop city** – Loses the key predictive variable entirely.

❌ **C. Region bucketing** – Reduces granularity, likely harming predictive power by grouping distinct cities.

❌ **D. Dataprep one-hot** – Creates hundreds or thousands of columns (one per city), introducing schema bloat and requiring manual file export—not minimal coding.

Question 19

You are an ML engineer at a bank that has a mobile app. Management wants biometric authentication via fingerprint. Fingerprints are highly sensitive PII and cannot be downloaded or stored. Which learning strategy should you recommend to train and deploy this ML model?

A. Data Loss Prevention API
B. Federated learning
C. MD5 to encrypt data
D. Differential privacy

✅ **Correct answer: B. Federated learning**

📌 Federated learning keeps raw fingerprint data on-device while training a global model by aggregating weight updates. No sensitive data leaves the device, preserving privacy and meeting compliance without central storage of PII.

Incorrect answers:

❌ **A. DLP API** – For detecting/shielding existing PII, not training a model on-device data.

❌ **C. MD5 encryption** – Irreversible hash; useless for training biometric models.

❌ **D. Differential privacy** – Adds noise to protect individual records but still requires central data aggregation; doesn't keep data fully on-device.

Question 20

You are an ML engineer in the contact center of a large enterprise. You need to build a sentiment analysis tool that predicts customer sentiment from recorded phone conversations. You must ensure gender, age, and cultural differences do not bias any stage of development or results. What should you do?

A. Convert the speech to text and extract sentiment based on sentences.
B. Convert the speech to text and build a model based on the words.
C. Extract sentiment directly from the voice recordings.
D. Convert the speech to text and extract sentiment using syntactical analysis.

✅ **Correct answer: C. Extract sentiment directly from the voice recordings.**
📌 Voice-based sentiment analysis (tone, pitch, prosody) bypasses text biases due to colloquialisms, gender/age speech patterns, or cultural language differences. By analyzing acoustic features, you avoid textual biases in ASR and NLP pipelines.

Incorrect answers:
❌ **A. Sentence-level text sentiment** – Relies on transcripts; subject to transcription errors and linguistic bias.
❌ **B. Word-level text model** – Similarly biased by vocabulary and translation nuances across demographics.
❌ **D. Syntactic text analysis** – Focuses on grammar and structure, still vulnerable to textual biases in language use.

Final Review Checklist & Exam Readiness Scorecard tailored for your Practice Exam

Before you schedule your exam, use these tools to ensure you're truly ready.

🗒️ Final Review Checklist

Use this checklist to validate that you're ready across all key exam domains:

🧠 Framing ML Problems & Business Requirements

☐ Can map business challenges to ML problem types (classification, regression, etc.)
☐ Understand ethical ML practices, stakeholder goals, and data governance needs
☐ Familiar with responsible AI principles, fairness, and explainability

🧪 Data Preparation & Feature Engineering

☐ Can clean, normalize, transform, and split datasets correctly
☐ Understand feature importance, dimensionality reduction, and encoding techniques
☐ Know how to use BigQuery, Dataflow, and Vertex AI Feature Store

⚙️ Model Development

☐ Understand when to use AutoML, custom training, and pre-trained models
☐ Comfortable with scikit-learn, TensorFlow, XGBoost, and BigQuery ML
☐ Know how to evaluate models (precision, recall, AUC, F1, confusion matrix)

🔁 ML Pipelines & Automation (MLOps)

☐ Can design training pipelines using Vertex AI Pipelines or Kubeflow
☐ Understand how to use Cloud Build, Artifact Registry, and CI/CD for ML workflows
☐ Familiar with model versioning, retraining triggers, and continuous delivery

📦 Deployment & Monitoring

☐ Know how to deploy models to Vertex AI Prediction (online + batch)
☐ Understand A/B testing, canary deployments, and rollback strategies
☐ Can monitor performance drift, model bias, and service uptime

🔐 Security, Compliance, and Cost Optimization

☐ Familiar with IAM roles for ML workloads and VPC-SC usage
☐ Understand data encryption at rest/in transit, DLP, and audit logging
☐ Can estimate costs using Vertex AI and optimize training/deployment pipelines

---

📊 Exam Readiness Scorecard

Rate your confidence per domain and list anything that needs review:

Domain	Confidence Level (1-5)	Notes / Gaps to Review
Framing ML Problems	⭐⭐⭐⭐☆	Revisit business metric alignment
Data Preparation & Feature Engineering	⭐⭐⭐⭐⭐	Fully confident
Model Development	⭐⭐⭐⭐☆	Review evaluation metrics tradeoffs
ML Pipelines & Automation	⭐⭐⭐☆☆	Need to rewatch Kubeflow pipeline demos
Deployment & Monitoring	⭐⭐⭐⭐☆	A/B rollout strategy review needed
Security & Cost Optimization	⭐⭐⭐⭐☆	Slightly unsure about VPC-SC scenarios

You're exam-ready when you're consistently hitting **4+ stars across all domains** and scoring **85%+ on full-length timed practice exams**.

💫 **Congratulations!! You are on the right path to certification, you made it to 20 questions so far. You're my kind of audience! BUT… you need all the questions to pass.**

Note: ***This is a courtesy PREVIEW DOC of around 20 questions. The complete practice exam with over 300 questions can be purchased at our Cloud Certification Store at*** *https://cloudcertificationstore.com/b/Y1OyW*

We personally took this exam recently, and quite a few of us, plus many of our buyers that leave a review, assure you, we had more than 90% of the same questions in the recent exam (read the reviews).

So go ahead, invest in your future, and find the complete exam for this particular certification, or more practice exams at our Cloud Certification Store at ***https://cloudcertificationstore.com/collection/all***

2025 Google Cloud Digital Leader Practice Exam Questions (GOOG-CDL-0010)

$9.00

NEW! Google Cloud Certified Generative AI Leader Practice Exam Questions (GOOG-GAIL-0010)

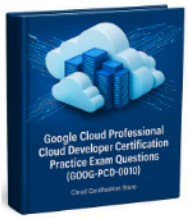
$9.00

2025 Google Cloud Associate Cloud Engineer Exam Questions (GOOG-ACE-0010)

$9.00

2025 Google Cloud Professional Cloud Architect Exam Questions (GOOG-PCA-0010)

$9.00



2025 Google Cloud Professional Cloud Developer Certification Practice Exam Questions (GOOG-PCD-0010)

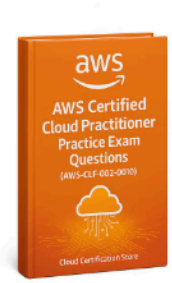
$9.00

2025 Google Cloud Professional Data Engineer Certification Practice Exam Questions (GOOG-PDE-0010)

$9.00

2025 Google Cloud Professional Cloud Database Engineer Practice Exam Questions (GOOG-PCDE-0010)

$9.00



2025 AWS Certified Cloud Practitioner Practice Exam Questions (AWS-CLF-002-0010)

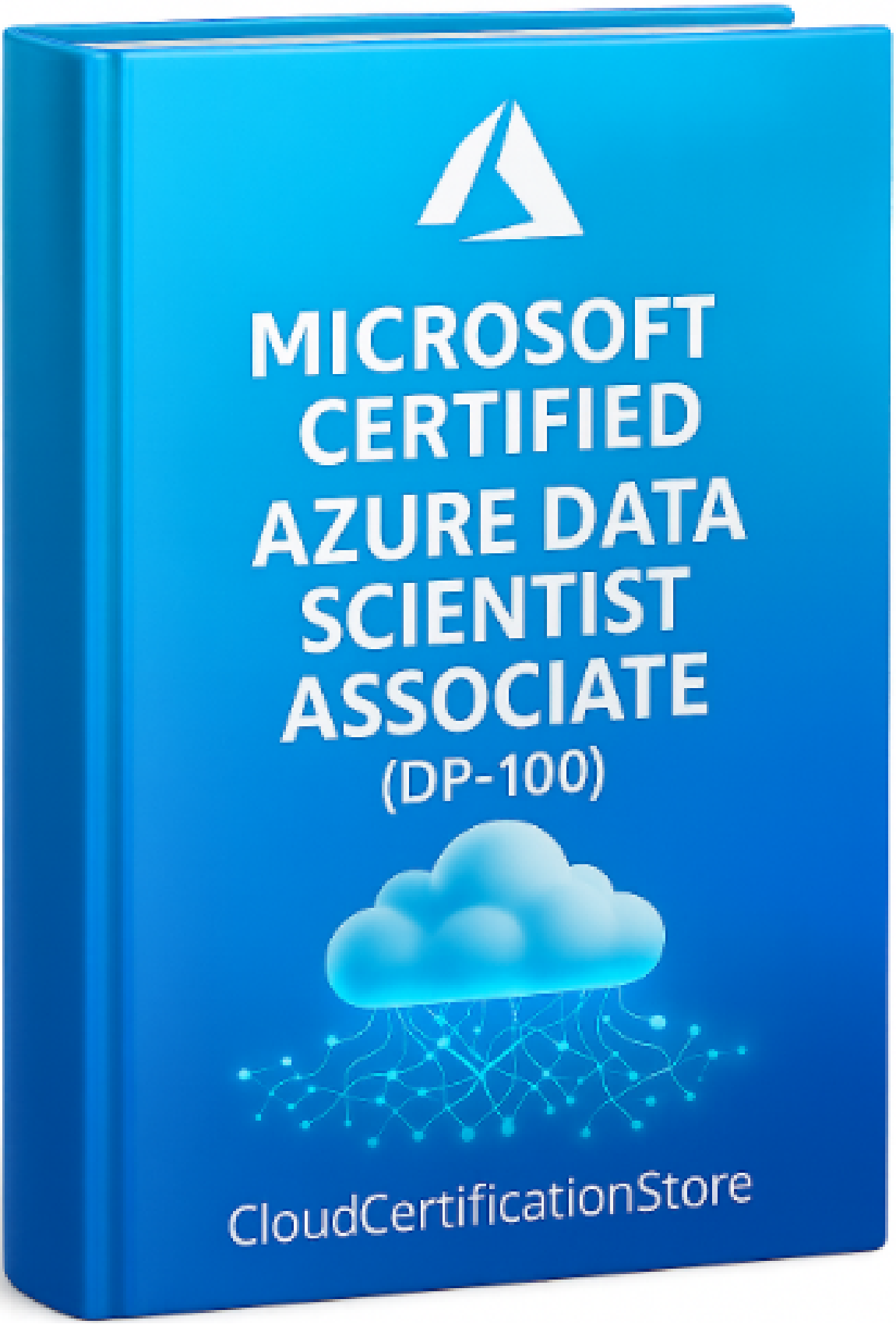
$9.00

2025 Google Cloud Professional Machine Learning Engineer Practice Exam Questions (GOOG-PMLE-0010)

$9.00

NEW! 2025 AWS Certified Solutions Architect - Associate SAA-C0300 Practice Exam Questions (AWS-SAA-C0300-0010)

$9.00

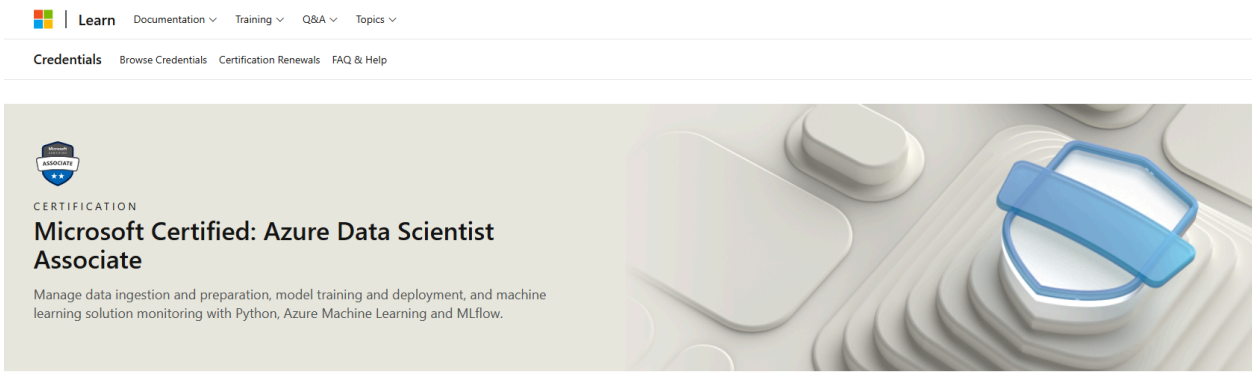

MICROSOFT
CERTIFIED
AZURE DATA
SCIENTIST
ASSOCIATE
(DP-100)
CloudCertificationStore

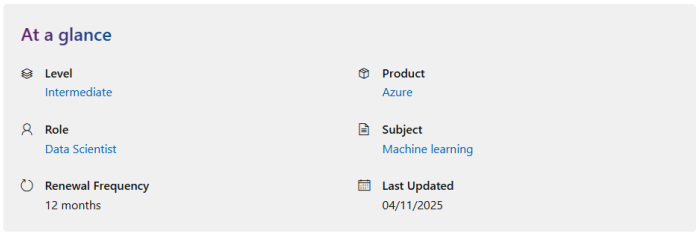
Microsoft Certified: Azure Data Scientist Associate (DP-100) - Practice Exam Questions (AZ-DP-100-0010)

DISCLAIMER

- This practice test includes questions **compiled from various exam preparation platforms.**
- Important: **Questions and answers were AI-assisted and human-curated**. Verify accuracy with official documentation before relying on this material.
- **Users are strongly encouraged to** double-check all content against **official documentation and trusted sources** before using it for exam preparation or making important decisions.
- The creators of this material assume **no responsibility** for any errors, inaccuracies, or outcomes, including exam results, based on the use of this content.
- **Some questions might be duplicated or close** to previous ones, this is done on purpose as a way to re-inforce your learning.
- Single-user licence only
    - Includes one unique Payhip Licence Key per purchase, along with a Product Key.
    - Redistribution, resale, or public posting is prohibited. We can trace any file to the purchaser, with the use of the purchased License Key and Product Key.

Microsoft Certified: Azure Data Scientist Associate (DP-100)





Overview

As a candidate for this certification, you should have subject matter expertise in applying data science and machine learning to implement and run machine learning workloads on Azure. Additionally, you should have knowledge of optimizing language models for AI applications using Azure AI.

Your responsibilities for this role include:

- Designing and creating a suitable working environment for data science workloads.
- Exploring data.
- Training machine learning models.
- Implementing pipelines.
- Running jobs to prepare for production.
- Managing, deploying, and monitoring scalable machine learning solutions.
- Using language models for building AI applications.

As a candidate for this certification, you should have knowledge and experience in data science by using:

- Azure Machine Learning
- MLflow
- Azure AI services, including Azure AI Search
- Azure AI Foundry

Skills earned upon completion

Prepare for the exam

Course
Designing and implementing a data science solution on Azure
Training in this course

Explore and configure the Azure Machine Learning workspace
3 hr 37 min
Learning Path
5 modules

Experiment with Azure Machine Learning
1 hr 10 min
Learning Path
2 modules

Optimize model training with Azure Machine Learning
2 hr 37 min
Learning Path
4 modules

Manage and review models in Azure Machine Learning
1 hr 22 min
Learning Path
2 modules

Deploy and consume models with Azure Machine Learning
1 hr 29 min
Learning Path
2 modules

Develop generative AI apps in Azure
7 hr 29 min
Learning Path
8 modules
Develop generative AI apps in Azur...

Practice for the exam

Practice Assessment

Assess your knowledge

Practice assessments provide you with an overview of the style, wording, and difficulty of the questions you're likely to experience on the exam. Through these assessments, you're able to assess your readiness, determine where additional preparation is needed, and fill knowledge gaps bringing you one step closer to the likelihood of passing your exam.

https://learn.microsoft.com/en-us/credentials/certifications/azure-data-scientist/practice/assessment?assessment-type=practice&assessmentId=62&practice-assessment-type=certification

Exam Sandbox

https://go.microsoft.com/fwlink/?linkid=2226877

Video

Exam DP-100 prep videos

Join our experts as they provide tips, tricks, and strategies for preparing for this Microsoft Certification exam.

https://learn.microsoft.com/en-us/shows/exam-readiness-zone/preparing-for-dp-100-design-and-prepare-a-machine-learning-solution-1-of-4

Take the exam

You will have 100 minutes to complete this assessment.

Exam policy

This exam will be proctored. You may have interactive components to complete as part of this exam. To learn more about exam duration and experience, visit: Exam duration and exam experience.

If you fail a certification exam, don't worry. You can retake it 24 hours after the first attempt. For subsequent retakes, the amount of time varies. For full details, visit: Exam retake policy.

Assessed on this exam

- Design and prepare a machine learning solution
- Explore data, and run experiments
- Train and deploy models
- Optimize language models for AI applications

Schedule the Exam

https://learn.microsoft.com/en-us/credentials/certifications/schedule-through-pearson-vue?examUid=exam.DP-100&examUrl=https://learn.microsoft.com/credentials/certifications

Price

$165 USD*

Price based on the country or region in which the exam is proctored.

Renew your certification

Do you know that Microsoft role-based and specialty certifications expire unless they are renewed? Learn the latest updates to the technology for your job role and renew your certification at no cost by passing an online assessment on Microsoft Learn.

Learn more about renewing

Certification resources

Exam DP-100 study guide

Focus your studies as you prepare for the exam. Review the study guide to learn about the topics the exam covers, updates, and additional resources.

Certification poster

Check out an overview of fundamentals, role-based, and specialty certifications.

Exam Replay

Boost your odds of success with this great offer.

Support for credentials

Get help through Microsoft Credentials support forums. A forum moderator will respond in one business day, Monday–Friday.

Choose your Microsoft Credential

Microsoft Applied Skills or Microsoft Certifications? Choose the path that fits your career goals, desired skillset, and schedule.

Study guide for Exam DP-100: Designing and Implementing a Data Science Solution on Azure

03/12/2025

Purpose of this document

This study guide should help you understand what to expect on the exam and includes a summary of the topics the exam might cover and links to additional resources. The information and materials in this document should help you focus your studies as you prepare for the exam.

Useful links	Description
How to earn the certification	Some certifications only require passing one exam, while others require passing multiple exams.
Certification renewal	Microsoft associate, expert, and specialty certifications expire annually. You can renew by passing a free online assessment on Microsoft Learn.
Your Microsoft Learn profile	Connecting your certification profile to Microsoft Learn allows you to schedule and renew exams and share and print certificates.

Exam scoring and score reports	A score of 700 or greater is required to pass.
Exam sandbox	You can explore the exam environment by visiting our exam sandbox.
Request accommodations	If you use assistive devices, require extra time, or need modification to any part of the exam experience, you can request an accommodation.
Take a free Practice Assessment	Test your skills with practice questions to help you prepare for the exam.

Updates to the exam

We always update the English language version of the exam first. Some exams are localized into other languages, and those are updated approximately eight weeks after the English version is updated. While Microsoft makes every effort to update localized versions as noted, there may be times when the localized versions of an exam are not updated on this schedule. Other available languages are listed in the Schedule Exam section of the Exam Details webpage. If the exam isn't available in your preferred language, you can request an additional 30 minutes to complete the exam.

Note

The bullets that follow each of the skills measured are intended to illustrate how we are assessing that skill. Related topics may be covered in the exam.

Note

Most questions cover features that are general availability (GA). The exam may contain questions on Preview features if those features are commonly used.

Skills measured as of April 11, 2025

Audience profile

As a candidate for this exam, you should have subject matter expertise in applying data science and machine learning to implement and run machine learning workloads on Azure. Additionally, you should have knowledge of optimizing language models for AI applications using Azure AI.

Your responsibilities for this role include:

- Designing and creating a suitable working environment for data science workloads.
- Exploring data.
- Training machine learning models.
- Implementing pipelines.
- Running jobs to prepare for production.
- Managing, deploying, and monitoring scalable machine learning solutions.
- Using language models for building AI applications.

As a candidate for this exam, you should have knowledge and experience in data science by using:

- Azure Machine Learning
- MLflow

- Azure AI services, including Azure AI Search
- Azure AI Foundry

Skills at a glance

- Design and prepare a machine learning solution (20–25%)
- Explore data, and run experiments (20–25%)
- Train and deploy models (25–30%)
- Optimize language models for AI applications (25–30%)

Design and prepare a machine learning solution (20–25%)

Design a machine learning solution

- Identify the structure and format for datasets
- Determine the compute specifications for machine learning workload
- Select the development approach to train a model

Create and manage resources in an Azure Machine Learning workspace

- Create and manage a workspace
- Create and manage datastores
- Create and manage compute targets
- Set up Git integration for source control

Create and manage assets in an Azure Machine Learning workspace

- Create and manage data assets
- Create and manage environments
- Share assets across workspaces by using registries

Explore data, and run experiments (20–25%)

Use automated machine learning to explore optimal models

- Use automated machine learning for tabular data
- Use automated machine learning for computer vision
- Use automated machine learning for natural language processing
- Select and understand training options, including preprocessing and algorithms
- Evaluate an automated machine learning run, including responsible AI guidelines

Use notebooks for custom model training

- Use the terminal to configure a compute instance
- Access and wrangle data in notebooks
- Wrangle data interactively with attached Synapse Spark pools and serverless Spark compute
- Retrieve features from a feature store to train a model
- Track model training by using MLflow
- Evaluate a model, including responsible AI guidelines

Automate hyperparameter tuning

- Select a sampling method
- Define the search space
- Define the primary metric
- Define early termination options

Train and deploy models (25–30%)

Run model training scripts

- Consume data in a job
- Configure compute for a job run

- Configure an environment for a job run
- Track model training with MLflow in a job run
- Define parameters for a job
- Run a script as a job
- Use logs to troubleshoot job run errors

Implement training pipelines

- Create custom components
- Create a pipeline
- Pass data between steps in a pipeline
- Run and schedule a pipeline
- Monitor and troubleshoot pipeline runs

Manage models

- Define the signature in the MLmodel file
- Package a feature retrieval specification with the model artifact
- Register an MLflow model
- Assess a model by using responsible AI principles

Deploy a model

- Configure settings for online deployment
- Deploy a model to an online endpoint
- Test an online deployed service
- Configure compute for a batch deployment
- Deploy a model to a batch endpoint
- Invoke the batch endpoint to start a batch scoring job

Optimize language models for AI applications (25–30%)

Prepare for model optimization

- Select and deploy a language model from the model catalog
- Compare language models using benchmarks
- Test a deployed language model in the playground
- Select an optimization approach

Optimize through prompt engineering and prompt flow

- Test prompts with manual evaluation
- Define and track prompt variants
- Create prompt templates
- Define chaining logic with the prompt flow SDK
- Use tracing to evaluate your flow

Optimize through Retrieval Augmented Generation (RAG)

- Prepare data for RAG, including cleaning, chunking, and embedding
- Configure a vector store
- Configure an Azure AI Search-based index store
- Evaluate your RAG solution

Optimize through fine-tuning

- Prepare data for fine-tuning
- Select an appropriate base model
- Run a fine-tuning job
- Evaluate your fine-tuned model

Study resources

We recommend that you train and get hands-on experience before you take the exam. We offer self-study options and classroom training as well as links to documentation, community sites, and videos.

Study resources	Links to learning and documentation
Get trained	Choose from self-paced learning paths and modules or take an instructor-led course
Find documentation	Azure Databricks Azure Machine Learning Azure Synapse Analytics MLflow and Azure Machine Learning
Ask a question	Microsoft Q&A | Microsoft Docs
Get community support	AI - Machine Learning - Microsoft Tech Community AI - Machine Learning Blog - Microsoft Tech Community
Follow Microsoft Learn	Microsoft Learn - Microsoft Tech Community
Find a video	Microsoft Learn Shows

Change log

The table below summarizes the changes between the current and previous version of the skills measured. The functional groups are in bold typeface followed by the objectives within each group. The table is a comparison between the previous and current version of the exam skills measured and the third column describes the extent of the changes.

Skill area prior to January 16, 2025	Skill area as of January 16, 2025	Change
Audience profile		Minor
Optimize language models for AI applications	Optimize language models for AI applications	No % change
Optimize through prompt engineering and Prompt flow	Optimize through prompt engineering and prompt flow	Minor

Additional resources

Documentation

Study guide for Exam AZ-500: Microsoft Azure Security Technologies

Study guide for Exam AZ-500: Microsoft Azure Security Technologies | Microsoft Docs

Study guide for Exam AZ-204: Developing Solutions for Microsoft Azure

Study guide for Exam AZ-204: Developing Solutions for Microsoft Azure | Microsoft Docs

Study guide for Exam AI-900: Microsoft Azure AI Fundamentals

Study guide for Exam AI-900: Microsoft Azure AI Fundamentals | Microsoft Docs

Practice Assessment

Practice Assessment

Training

Learning path

Solution Architect: Design Microsoft Power Platform solutions - Training

Learn how a solution architect designs solutions.

Certification

Microsoft Certified: Azure Data Scientist Associate - Certifications

Manage data ingestion and preparation, model training and deployment, and machine learning solution monitoring with Python, Azure Machine Learning and MLflow.

Practice Questions - PREVIEW 20 out of 500+

Question 1

You must train a machine learning model in Azure Machine Learning using a large dataset stored in Azure Blob Storage. The goal is to minimize data-copy overhead while enabling distributed training with dynamic scaling.

A. Compute Instance
B. Compute Cluster
C. Attached Compute to Synapse Spark Pool
D. Kubernetes Service

✅ **Correct Answer: C. Attached Compute to Synapse Spark Pool**

📌 Using an attached Synapse Spark pool provides serverless distributed compute, allowing direct access to data in Blob Storage and dynamic scaling for big-data workloads. It eliminates unnecessary data movement and is fully integrated with Azure ML pipelines.

❌ **Why not the other options?**
A. Compute Instance: A single-node environment ideal for experimentation, not for distributed or large-scale training.
B. Compute Cluster: Suitable for parallel training but still requires data copy into compute nodes, adding overhead.
D. Kubernetes Service: Used primarily for deployment and inference, not for distributed training integration with Blob datasets.

Question 2

You need to evaluate a binary classification model and use precision as the primary metric. Which visualization should you choose in Azure Machine Learning Studio?

A. ROC Curve
B. Confusion Matrix
C. Lift Chart
D. Venn Diagram

✅ **Correct Answer: B. Confusion Matrix**

📌 The confusion matrix provides direct counts of true positives, false positives, and false negatives, which allows for easy computation of precision and recall. It is the standard visualization for assessing classification performance metrics.

❌ **Why not the other options?**
A. ROC Curve: Measures trade-offs between true and false positive rates, not precision specifically.
C. Lift Chart: Evaluates ranking performance in classification, not direct precision calculation.
D. Venn Diagram: A conceptual visualization, not used for quantitative metric evaluation in ML workflows.

Question 3

You plan to train a speech-recognition deep learning model on a Data Science Virtual Machine (DSVM). The model must support GPU acceleration and the latest Python version.

A. Theano
B. TensorFlow
C. Scikit-learn
D. CNTK

✅ **Correct Answer: B. TensorFlow**

📌 TensorFlow provides native CUDA support for GPUs and full compatibility with the latest Python versions, making it ideal for speech recognition models requiring deep learning on DSVM.

❌ **Why not the other options?**
A. Theano: Deprecated and no longer actively supported by Azure environments.
C. Scikit-learn: Designed for traditional ML, not for large-scale GPU-based deep learning.
D. CNTK: While Microsoft-developed, it's discontinued in favor of TensorFlow and PyTorch integration.

Question 4

A convolutional neural network (CNN) for image classification is showing signs of overfitting. You need to improve model generalization.

A. Reduce training data size
B. Add L1/L2 regularization and apply data augmentation
C. Add more dense layers
D. Remove dropout

✅ **Correct Answer: B. Add L1/L2 regularization and apply data augmentation**

📌 L1/L2 regularization penalizes large weights, while data augmentation generates diverse samples, both reducing overfitting and improving the model's ability to generalize unseen data.

❌ **Why not the other options?**
A. Reduce training data size: This would worsen overfitting by limiting exposure to variation.
C. Add more dense layers: Increases complexity, exacerbating overfitting.
D. Remove dropout: Dropout is used to combat overfitting; removing it increases model variance.

Question 5

In Azure Machine Learning Designer, you must split a dataset into training and testing subsets. Which module should you use?

A. Group Categorical Values
B. Split Data
C. Clip Values
D. Edit Metadata

✅ **Correct Answer: B. Split Data**

📌 The Split Data module is designed to partition datasets for training and testing, ensuring balanced sampling and reproducibility in ML experiments.

❌ **Why not the other options?**
A. Group Categorical Values: Used for merging string categories, not data partitioning.
C. Clip Values: Used to restrict numeric ranges, unrelated to data splitting.
D. Edit Metadata: Adjusts data types and field roles, not used for partitioning.

Question 6

You configure k-fold cross-validation for a dataset and must select the most common k value.

A. 3
B. 5 or 10
C. 1
D. 100

✅ **Correct Answer: B. 5 or 10**

📌 Using k=5 or k=10 provides a balance between computational cost and robust validation, minimizing bias and variance in model evaluation.

❌ **Why not the other options?**
A. 3: Provides weaker statistical confidence in validation.
C. 1: Equivalent to no cross-validation.
D. 100: Computationally excessive and unnecessary for most datasets.

Question 7

You use the Clean Missing Data module and choose "Replace with Median." When is this method appropriate?

A. For categorical features
B. For continuous numeric features
C. For text features
D. Never

✅ **Correct Answer: B. For continuous numeric features**

📌 The median is a robust central tendency measure for numeric data and helps handle outliers while imputing missing values effectively.

❌ **Why not the other options?**
A. Categorical features: Should use mode or most frequent value instead.
C. Text features: Text cannot be imputed numerically.
D. Never: Median replacement is a standard numeric imputation technique.

Question 8

When configuring hyperparameter tuning in Azure ML, you want to iterate efficiently through parameter combinations with reduced compute cost. Which sweep mode should you select?

A. Measured grid
B. Entire grid
C. Random grid
D. Selective grid

✅ **Correct Answer: C. Random grid**

📌 Random grid sweep randomly samples from parameter space, achieving high coverage with lower compute demand than exhaustive grid search.

❌ **Why not the other options?**
A. Measured grid: Not a valid tuning mode.
B. Entire grid: Tests all combinations, increasing cost.
D. Selective grid: Not available as a sweep configuration.

Question 9

You must import a CSV dataset from a public web URL into Azure ML Designer with the least administrative effort.

A. Import Data
B. Dataset
C. Copy Data
D. Convert to TXT

✅ **Correct Answer: A. Import Data**

📌 The Import Data module enables direct ingestion from public URLs or cloud storage with no registration or prior setup, ideal for quick experiments.

❌ **Why not the other options?**
B. Dataset: Requires manual dataset registration.
C. Copy Data: Used for ETL pipelines, not ingestion in Designer.
D. Convert to TXT: Only changes file format, doesn't import data.

Question 10

Which compute target allows you to directly connect Azure Machine Learning Designer to a Spark-based big data environment for model training?

A. Compute Instance
B. Compute Cluster
C. Attached Synapse Spark Pool
D. Inference Cluster

✅ **Correct Answer: C. Attached Synapse Spark Pool**

📌 An attached Synapse Spark Pool provides seamless Spark integration, enabling large-scale distributed data processing directly from Azure ML Designer.

❌ **Why not the other options?**
A. Compute Instance: Local compute, not distributed.

B. Compute Cluster: Suitable for ML jobs but not Spark integration.
D. Inference Cluster: Used for deployment, not training.

Question 11

You need to transform categorical features into binary columns suitable for model training. Which module should you use?

A. Clean Missing Data
B. Convert to Indicator Values
C. Edit Metadata
D. Group Categorical Values

✅ **Correct Answer: B. Convert to Indicator Values**

📌 The Convert to Indicator Values module performs one-hot encoding, converting categorical fields into binary numeric columns for ML algorithms.

❌ **Why not the other options?**
A. Clean Missing Data: Handles nulls, not encoding.
C. Edit Metadata: Alters data types, not values.
D. Group Categorical Values: Used for combining similar labels.

Question 12

You must visually identify outliers in a dataset. Which method provides this capability?

A. ROC Curve
B. Box Plot
C. Histogram
D. Line Chart

✅ **Correct Answer: B. Box Plot**

📌 A box plot visualizes the data distribution and highlights outliers through whisker boundaries, making it ideal for anomaly inspection.

❌ **Why not the other options?**
A. ROC Curve: Evaluates model performance, not raw data.
C. Histogram: Shows frequency but not outlier thresholds.
D. Line Chart: Displays trends, not statistical deviations.

Question 13

In a Python experiment, you need to record a metric for the number of rows processed. Which statement logs this metric in Azure ML?

A. `run.upload_file('row_count', './data.csv')`
B. `run.log('row_count', rows)`
C. `run.log_row('row_count', rows)`
D. `run.tag('row_count', rows)`

✅ **Correct Answer: B. run.log('row_count', rows)**

📌 The `run.log()` method records scalar values as metrics, retrievable after the experiment completes for tracking performance.

❌ **Why not the other options?**
A. upload_file: Uploads artifacts, not metrics.
C. log_row: Used for tabular data, not single values.
D. tag: Assigns labels, not numeric logs.

Question 14

You are developing a text translation model that must learn language sequences. Which architecture should you choose?

A. CNN
B. RNN
C. GAN
D. Transformer

✅ **Correct Answer: B. RNN**

📌 RNNs capture temporal dependencies in sequential data such as language, enabling context-aware translation tasks.

❌ **Why not the other options?**
A. CNN: Processes spatial data (images).
C. GAN: Used for generative image synthesis.
D. Transformer: Modern alternative but not explicitly required in this scenario.

Question 15

You must increase underrepresented class samples in your dataset. Which Azure ML module should you use?

A. Join Data
B. SMOTE
C. Group Categorical Values
D. Clip Values

✅ **Correct Answer: B. SMOTE**

📌 SMOTE (Synthetic Minority Oversampling Technique) generates synthetic minority samples, improving class balance for better classification.

❌ **Why not the other options?**
A. Join Data: Used for merging datasets.
C. Group Categorical Values: Combines categories, not records.
D. Clip Values: Used for numeric scaling, not sampling.

Question 16

You create a batch inference pipeline using Azure ML SDK. Where will the output file be located?

A. Activity Log
B. digit_identification.py
C. parallel_run_step.txt in output folder
D. Inference Clusters tab

✅ **Correct Answer: C. parallel_run_step.txt in output folder**

📌 Batch pipelines aggregate outputs into a single text file named `parallel_run_step.txt`, stored in the designated output directory.

❌ **Why not the other options?**
A. Activity Log: Tracks events, not outputs.
B. digit_identification.py: Script source file, not result storage.
D. Inference Clusters: Used for deployment endpoints, not batch logs.

Question 17

During AutoML classification, you must restrict experiments to linear algorithms. What should you configure?

A. Enable automatic featurization
B. Disable deep learning
C. Turn off featurization
D. Enable forecasting

✅ **Correct Answer: B. Disable deep learning**

📌 Disabling deep learning ensures AutoML evaluates only classical linear algorithms, such as logistic regression or linear SVMs.

❌ **Why not the other options?**
A. Enable automatic featurization: Influences preprocessing, not algorithm choice.
C. Turn off featurization: Reduces model accuracy, not algorithm control.
D. Forecasting: Changes task type, not model selection.

Question 18

When registering an MLflow model in Azure ML, which file defines its signature and dependencies?

A. model.pkl
B. MLmodel
C. conda.yaml
D. run.json

✅ **Correct Answer: B. MLmodel**

📌 The MLmodel file specifies the model's entry points, environment, and metadata for deployment consistency across platforms.

❌ **Why not the other options?**
A. model.pkl: Stores serialized weights, not metadata.
C. conda.yaml: Lists environment dependencies only.
D. run.json: Contains execution logs, not model definitions.

Question 19

You must select a machine learning environment that supports disconnected training using Caffe2 or Chainer frameworks on personal devices.

A. Azure ML Service with DSVM
B. Azure ML Studio
C. Azure Databricks
D. Azure Kubernetes Service

✅ **Correct Answer: A. Azure ML Service with DSVM**

📌 DSVM supports both Caffe2 and Chainer frameworks locally and can synchronize pipelines to Azure ML when reconnected.

❌ **Why not the other options?**
B. ML Studio: Requires full connectivity.
C. Databricks: Cluster-based, not offline.
D. AKS: For deployment, not training.

Question 20

You are optimizing a Retrieval-Augmented Generation (RAG) workflow. Which two Azure components are mandatory?

A. Azure AI Search index store and vector store
B. Azure Key Vault and Azure Monitor
C. Compute Cluster and Batch Endpoint
D. Prompt Flow and Pipeline Component

✅ **Correct Answer: A. Azure AI Search index store and vector store**

📌 RAG relies on a vector store for embeddings and an Azure AI Search index for retrieval, allowing LLMs to ground responses in enterprise data.

❌ **Why not the other options?**
B. Key Vault & Monitor: Handle secrets and metrics, not retrieval.
C. Compute Cluster & Batch Endpoint: Manage compute, not RAG data.
D. Prompt Flow: Useful for orchestration, but not required for retrieval storage.

(END OF PREVIEW QUESTIONS)

Final Review Checklist & Exam Readiness Scorecard

✅ How to Use the Final Review Checklist

This section is meant to **validate your hands-on skills and theoretical readiness** across all exam topics.

Step-by-step:

1. **Print it or load it in a note-taking app** (Notion, Google Docs, OneNote, etc.).
2. Go through each checkbox:
    - ✅ Check it if you **fully understand and can implement** the topic without looking up documentation.
    - ❌ Leave it unchecked if you feel unsure or haven't practiced the task.
3. Prioritize unchecked topics by reviewing:
    - Check the official documentation
    - Practice exams
    - Hands-on labs
4. For each **unchecked item**, write a short action plan or resource link next to it.

📈 How to Use the Exam Readiness Scorecard

This part helps you **self-assess your confidence level** and **focus your revision time** wisely.

Instructions:

1. For each domain (e.g., "Hybrid connectivity and routing"), **rate yourself** from 1 to 5:
    - 1️⃣ = No understanding or hands-on practice
    - 3️⃣ = Moderate familiarity, but need review
    - 5️⃣ = Mastered topic and can apply it in real-world use
2. Add **Notes / Action Items** to explain:
    - Why you scored yourself low
    - What resources you'll use to improve (YouTube, whitepapers, exam guides)
    - Practice test scores if relevant
3. Reassess **2–3 days before your exam**, and compare scores to measure improvement.

🧠 Bonus Tips

- Do **timed mock exams** and cross-reference errors with checklist topics
- Use the scorecard to **simulate an exam debrief**: where did you fail? What must you strengthen?

Once all checklist items are ✅ and all categories are at **4–5 stars** and you're consistently scoring **85%+** on full practice exams with confidence in scenario-based reasoning, then 🎯 you're likely **ready to book the real exam.**

Final Review Checklist

🧩 Design and Prepare a Machine Learning Solution (20 – 25%)

☐ Identify dataset structures, formats, and compute requirements

☐ Choose appropriate ML algorithms and training approaches

☐ Create and manage Azure Machine Learning workspaces

☐ Create and manage datastores and compute targets

☐ Configure Git integration for version control

☐ Create and manage data assets and environments

☐ Share assets across workspaces using registries

🔍 Explore Data and Run Experiments (20 – 25%)

☐ Use Automated ML for tabular, computer-vision, and NLP tasks

☐ Evaluate AutoML runs with Responsible AI guidelines

☐ Use notebooks to access and wrangle data interactively

☐ Attach and use Synapse Spark or serverless Spark compute

☐ Retrieve features from Feature Store to train models

☐ Track experiments and metrics with MLflow

☐ Evaluate trained models following Responsible AI principles

☐ Perform hyperparameter tuning (sampling, search space, primary metric, early termination)

🧠 Train and Deploy Models (25 – 30%)

☐ Run training scripts and jobs, define parameters, and troubleshoot logs

☐ Configure compute environments for job runs

☐ Implement custom components and pipelines

☐ Pass data between pipeline steps and schedule runs

☐ Monitor and troubleshoot pipeline execution

☐ Register and manage MLflow models

☐ Define model signatures and package feature retrieval specs

☐ Deploy models to online and batch endpoints

☐ Test, invoke, and monitor deployed services

⚙️ Optimize Language Models for AI Applications (25 – 30%)

☐ Select, deploy, and benchmark language models from the catalog

☐ Test deployed LLMs in the playground and compare results

☐ Apply optimization via prompt engineering and Prompt Flow

☐ Design prompt templates, track variants, and define chaining logic

☐ Use tracing for evaluation of prompt flows

☐ Implement Retrieval-Augmented Generation (RAG): cleaning, chunking, embedding

☐ Configure vector stores and Azure AI Search-based indexes

☐ Evaluate RAG solutions for accuracy and latency

☐ Perform fine-tuning: prepare data, choose base model, run jobs, evaluate results

Exam Readiness Scorecard

Domain	Confidence (1–5)	Notes / Action Items
🧩 Design & Prepare ML Solutions (20–25%)	☐1 ☐2 ☐3 ☐4 ☐5	
🔍 Explore Data & Run Experiments (20–25%)	☐1 ☐2 ☐3 ☐4 ☐5	
🧠 Train & Deploy Models (25–30%)	☐1 ☐2 ☐3 ☐4 ☐5	
⚙️ Optimize Language Models for AI (25–30%)	☐1 ☐2 ☐3 ☐4 ☐5	
⏱ Time Management (150-min exam pacing)	☐1 ☐2 ☐3 ☐4 ☐5	Complete 50-question timed set practice

🧭 How to Use

1. Check off each line once you can *perform the task hands-on in Azure ML Studio or SDK*.
2. Rate each domain (1 = weak → 5 = expert) to focus your revision.
3. Prioritize high-weight domains (**Training & Deployment**, **LLM Optimization**) in your final review week.

4. Re-test using Microsoft's **Practice Assessment** and **sandbox** to simulate the exam interface.

📝 Tips for Use

- Start by marking each checklist item you feel confident about; leave blanks for uncertain areas.
- For each blank or lower-confidence domain, gather relevant Microsoft Learn modules, hands-on labs, or practice problems to reinforce.
- Use timed mock exams to simulate pacing and identify weak spots in your scorecard.
- Revisit the checklist closer to your exam and ensure all critical areas are covered.

🟢 You're ready when all sections score **4 or 5 stars**, and your practice test scores are consistently **above 85%**.

💫 **Congratulations!! You are on the right path to certification.**

All of our practice exams include **300+ questions**. This one has **over 500.**

Get the full version here https://cloudcertificationstore.com/b/hvask

Our writers who have taken the exam recently—and the reviewers who purchased these materials—agree that **over 90 %** of the questions matched what they saw on the live test.

Invest in your future: browse the full catalogue of Cloud practice exams at our store

Featured Collection

2025 Google Cloud Professional Cloud Architect Exam Questions (GOOG-PCA-0010)

~~$8.89~~ $6.22

2025 Google Cloud Professional Cloud Developer Certification Practice Exam Questions (GOOG-PCD-0010)

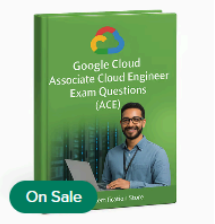
~~$8.89~~ $6.22

2025 Google Cloud Professional Machine Learning Engineer Practice Exam Questions (GOOG-PMLE-0010)

~~$8.99~~ $6.29

2025 Google Cloud Professional Data Engineer Certification Practice Exam Questions (GOOG-PDE-0010)

~~$8.89~~ $6.22

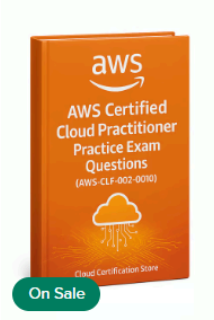


2025 Google Cloud Associate Cloud Engineer Exam Questions (GOOG-ACE-0010)

~~$8.89~~ $6.22

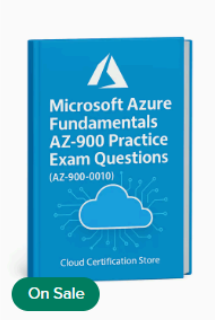
2025 Google Cloud Professional Cloud Database Engineer Practice Exam Questions (GOOG-PCDE-0010)

~~$8.89~~ $6.22



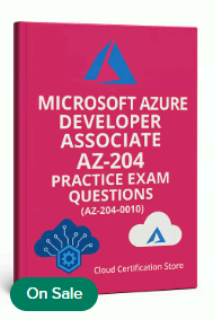
2025 AWS Certified Cloud Practitioner Practice Exam Questions (AWS-CLF-002-0010)

~~$8.99~~ $6.29



2025 Microsoft Azure Fundamentals AZ-900 Practice Exam Questions (AZ-900-0010)

~~$8.89~~ $6.22



2025 Microsoft Azure Developer AZ-204 Practice Exam Questions (AZ-204-0010)

~~$8.89~~ $6.22

FREE Practice Questions for the Google Cloud Generative AI Leader Exam

$0.00+