Free Amazon MLA-C01 Practice Questions

Shared by Norton on 12-08-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

A company is building an Amazon SageMaker AI pipeline for an ML model. The pipeline uses distributed processing and training.

An ML engineer needs to encrypt network communication between instances that run distributed jobs. The ML engineer configures the distributed jobs to run in a private VPC.

What should the ML engineer do to meet the encryption requirement?

Options:

A- Enable network isolation.
B- Configure traffic encryption by using security groups.
C- Enable inter-container traffic encryption.
D- Enable VPC flow logs.

Answer:

C

Explanation:

In distributed training and processing jobs, multiple instances and containers communicate with each other over the network to exchange gradients, parameters, and intermediate results. Even when jobs run inside a private VPC, network traffic between instances is not automatically encrypted at the application layer.

AWS documentation for Amazon SageMaker specifies that inter-container traffic encryption is the supported mechanism for encrypting data in transit between containers that participate in distributed training or processing jobs. When enabled, SageMaker uses TLS to encrypt all communication between containers across instances, ensuring confidentiality and compliance with security requirements.

Option A (network isolation) prevents containers from making outbound network calls but does not encrypt traffic between distributed instances. Option B is incorrect because security groups control traffic access, not encryption. Option D (VPC flow logs) is a monitoring feature and does not provide encryption.

Therefore, enabling inter-container traffic encryption is the correct and AWS-recommended solution.

Question 2

Question Type: MultipleChoice

A company is building a near real-time data analytics application to detect anomalies and failures for industrial equipment. The company has thousands of IoT sensors that send data every 60 seconds. When new versions of the application are released, the company wants to ensure that application code bugs do not prevent the application from running.

Which solution will meet these requirements?

Options:

A- Use Amazon Managed Service for Apache Flink with the system rollback capability enabled to build the data analytics application.
B- Use Amazon Managed Service for Apache Flink with manual rollback when an error occurs to build the data analytics application.
C- Use Amazon Data Firehose to deliver real-time streaming data programmatically for the data analytics application. Pause the stream when a new version of the application is released and resume the stream after the application is deployed.
D- Use Amazon Data Firehose to deliver data to Amazon EC2 instances across two Availability Zones for the data analytics application.

Answer:

A

Explanation:

For near real-time anomaly detection on streaming IoT data, AWS recommends Amazon Managed Service for Apache Flink. Flink is designed for stateful, low-latency stream processing and is well suited for time-series sensor analytics.

A key requirement is application resilience during deployments. Managed Flink supports system rollback, which automatically reverts to the last stable application version if a new deployment fails. This capability ensures uninterrupted processing even if application bugs are introduced, meeting the company's reliability requirement.

Manual rollback (Option B) introduces operational risk and delays. Amazon Data Firehose (Options C and D) is a delivery service and does not support complex anomaly detection logic or application version rollback.

Therefore, using Managed Flink with system rollback enabled is the correct solution.

Question 3

Question Type: MultipleChoice

An ML engineer wants to run a training job on Amazon SageMaker AI by using multiple GPUs. The training dataset is stored in Apache Parquet format.

The Parquet files are too large to fit into the memory of the SageMaker AI training instances.

Which solution will fix the memory problem?

Options:

A- Attach an Amazon EBS Provisioned IOPS SSD volume and store the files on the EBS volume.
B- Repartition the Parquet files by using Apache Spark on Amazon EMR and use the repartitioned files for training.
C- Change to memory-optimized instance types with sufficient memory.
D- Use SageMaker distributed data parallelism (SMDDP) to split memory usage.

Answer:

B

Explanation:

Large Parquet files can cause out-of-memory (OOM) issues during training if individual files exceed the memory capacity of the training instances. AWS documentation recommends repartitioning large datasets into smaller files to enable efficient streaming and parallel loading.

By using Apache Spark on Amazon EMR to repartition the Parquet files, the ML engineer can split the dataset into multiple smaller files that can be read incrementally during training. This approach avoids loading a single large file into memory and improves data parallelism.

Attaching larger EBS volumes increases storage capacity but does not solve memory constraints. Switching to memory-optimized instances increases cost and is not necessary when the dataset can be restructured. SageMaker distributed data parallelism focuses on model parameter synchronization across GPUs, not dataset file size.

AWS best practices explicitly recommend partitioning large Parquet datasets to improve memory efficiency during training.

Therefore, Option B is the correct and AWS-aligned solution.

Question 4

Question Type: MultipleChoice

A digital media entertainment company needs real-time video content moderation to ensure compliance during live streaming events.

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

Options:

A- Use Amazon Rekognition and AWS Lambda to extract and analyze the metadata from the videos' image frames.
B- Use Amazon Rekognition and a large language model (LLM) hosted on Amazon Bedrock to extract and analyze the metadata from the videos' image frames.
C- Use Amazon SageMaker AI to extract and analyze the metadata from the videos' image frames.
D- Use Amazon Transcribe and Amazon Comprehend to extract and analyze the metadata from the videos' image frames.

Answer:

A

Explanation:

For real-time video content moderation with minimal operational overhead, AWS documentation recommends using fully managed, purpose-built AI services. Amazon Rekognition provides real-time video analysis capabilities, including content moderation, unsafe content detection, and label recognition for live video streams.

By integrating Rekognition with AWS Lambda, the company can automatically process video frames, extract moderation metadata, and take immediate action (such as flagging or stopping a stream) without managing servers, models, or infrastructure. This serverless architecture scales automatically and minimizes operational complexity.

Option B introduces unnecessary complexity. While Amazon Bedrock LLMs are powerful, they are not required for image-based moderation tasks that Rekognition already handles natively.

Option C is incorrect because using Amazon SageMaker would require model training, endpoint management, and scaling, significantly increasing operational overhead.

Option D is incorrect because Amazon Transcribe and Amazon Comprehend are designed for audio and text analysis, not image or video frame moderation.

Therefore, Amazon Rekognition with AWS Lambda is the most efficient, scalable, and low-maintenance solution for real-time video moderation during live streaming events.

Question 5

Question Type: MultipleChoice

A company has developed a new ML model. The company requires online model validation on 10% of the traffic before the company fully releases the model in production. The company uses an Amazon SageMaker endpoint behind an Application Load Balancer (ALB) to serve the model.

Which solution will set up the required online validation with the LEAST operational overhead?

Options:

A- Use production variants to add the new model to the existing SageMaker endpoint. Set the variant weight to 0.1 for the new model. Monitor the number of invocations by using Amazon CloudWatch.
B- Use production variants to add the new model to the existing SageMaker endpoint. Set the variant weight to 1 for the new model. Monitor the number of invocations by using Amazon CloudWatch.
C- Create a new SageMaker endpoint. Use production variants to add the new model to the new endpoint. Monitor the number of invocations by using Amazon CloudWatch.
D- Configure the ALB to route 10% of the traffic to the new model at the existing SageMaker endpoint. Monitor the number of invocations by using AWS CloudTrail.

Answer:

A

Explanation:

Scenario: The company wants to perform online validation of a new ML model on 10% of the traffic before fully deploying the model in production. The setup must have minimal operational overhead.

Why Use SageMaker Production Variants?

Built-In Traffic Splitting: Amazon SageMaker endpoints support production variants, allowing multiple models to run on a single endpoint. You can direct a percentage of incoming traffic to each variant by adjusting the variant weights.

Ease of Management: Using production variants eliminates the need for additional infrastructure like separate endpoints or custom ALB configurations.

Monitoring with CloudWatch: SageMaker automatically integrates with CloudWatch, enabling real-time monitoring of model performance and invocation metrics.

Steps to Implement:

Deploy the New Model as a Production Variant:

Update the existing SageMaker endpoint to include the new model as a production variant. This can be done via the SageMaker console, CLI, or SDK.

Example SDK Code:

```
import boto3

sm_client = boto3.client('sagemaker')

response = sm_client.update_endpoint_weights_and_capacities(

EndpointName='existing-endpoint-name',

DesiredWeightsAndCapacities=[

{'VariantName': 'current-model', 'DesiredWeight': 0.9},

{'VariantName': 'new-model', 'DesiredWeight': 0.1}

]

)
```

Set the Variant Weight:

Assign a weight of 0.1 to the new model and 0.9 to the existing model. This ensures 10% of traffic goes to the new model while the remaining 90% continues to use the current model.

Monitor the Performance:

Use Amazon CloudWatch metrics, such as InvocationCount and ModelLatency, to monitor the traffic and performance of each variant.

Validate the Results:

Analyze the performance of the new model based on metrics like accuracy, latency, and failure rates.

Why Not the Other Options?

Option B: Setting the weight to 1 directs all traffic to the new model, which does not meet the requirement of splitting traffic for validation.

Option C: Creating a new endpoint introduces additional operational overhead for traffic routing and monitoring, which is unnecessary given SageMaker's built-in production variant capability.

Option D: Configuring the ALB to route traffic requires manual setup and lacks SageMaker's seamless variant monitoring and traffic splitting features.

Conclusion:

Using production variants with a weight of 0.1 for the new model on the existing SageMaker endpoint provides the required traffic split for online validation with minimal operational overhead.

Amazon SageMaker Endpoints

SageMaker Production Variants

Monitoring SageMaker Endpoints with CloudWatch

Question 6

Question Type: MultipleChoice

An ML engineer needs to use an ML model to predict the price of apartments in a specific location.

Which metric should the ML engineer use to evaluate the model's performance?

Options:

A- Accuracy
B- Area Under the ROC Curve (AUC)
C- F1 score
D- Mean absolute error (MAE)

Answer:

D

Explanation:

When predicting continuous variables, such as apartment prices, it's essential to evaluate the model's performance using appropriate regression metrics. The Mean Absolute Error (MAE) is a widely used metric for this purpose.

Understanding Mean Absolute Error (MAE):

MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It calculates the average absolute difference between predicted values and actual values, providing a straightforward interpretation of prediction accuracy.

Formula: $\mathrm{MAE} = \frac{1}{n}\sum_{i-1}^{n} |y_i - \hat{y}_i|$ Where:

- n = Number of data points
- y_i = Actual value
- $\hat{y}_i$ = Predicted value

Advantages of MAE:

Interpretability: MAE is expressed in the same units as the target variable, making it easy to understand.

Robustness to Outliers: Unlike metrics that square the errors (e.g., Mean Squared Error), MAE does not disproportionately penalize larger errors, making it more robust to outliers.

Comparison with Other Metrics:

Accuracy, AUC, F1 Score: These metrics are designed for classification tasks, where the goal is to predict discrete labels. They are not suitable for regression problems involving continuous target variables.

Mean Squared Error (MSE): While MSE also measures prediction errors, it squares the differences, giving more weight to larger errors. This can be useful in certain contexts but may be sensitive to outliers.

Conclusion:

For evaluating the performance of a model predicting apartment prices---a continuous variable---MAE is an appropriate and effective metric. It provides a clear indication of the average prediction error in the same units as the target variable, facilitating straightforward interpretation and comparison.

Regression Metrics -- GeeksforGeeks

Evaluation Metrics for Your Regression Model -- Analytics Vidhya

Regression Metrics for Machine Learning -- Machine Learning Mastery

Question 7

Question Type: MultipleChoice

Case Study

A company is building a web-based AI application by using Amazon SageMaker. The application will provide the following capabilities and features: ML experimentation, training, a

central model registry, model deployment, and model monitoring.

The application must ensure secure and isolated use of training data during the ML lifecycle. The training data is stored in Amazon S3.

The company needs to run an on-demand workflow to monitor bias drift for models that are deployed to real-time endpoints from the application.

Which action will meet this requirement?

Options:

A- Configure the application to invoke an AWS Lambda function that runs a SageMaker Clarify job.
B- Invoke an AWS Lambda function to pull the sagemaker-model-monitor-analyzer built-in SageMaker image.
C- Use AWS Glue Data Quality to monitor bias.
D- Use SageMaker notebooks to compare the bias.

Answer:

A

Explanation:

Monitoring bias drift in deployed machine learning models is crucial to ensure fairness and accuracy over time. Amazon SageMaker Clarify provides tools to detect bias in ML models, both during training and after deployment. To monitor bias drift for models deployed to real-time endpoints, an effective approach involves orchestrating SageMaker Clarify jobs using AWS Lambda functions.

Implementation Steps:

Set Up Data Capture:

Enable data capture on the SageMaker endpoint to record input data and model predictions. This captured data serves as the basis for bias analysis.

Develop a Lambda Function:

Create an AWS Lambda function configured to initiate a SageMaker Clarify job. This function will process the captured data to assess bias metrics.

Schedule or Trigger the Lambda Function:

Configure the Lambda function to run on-demand or at scheduled intervals using Amazon CloudWatch Events or EventBridge. This setup allows for regular bias monitoring as per the application's requirements.

Analyze and Respond to Results:

After each Clarify job completes, review the generated bias reports. If bias drift is detected, take appropriate actions, such as retraining the model or adjusting data preprocessing steps.

Advantages of This Approach:

Automation: Utilizing AWS Lambda for orchestrating Clarify jobs enables automated and scalable bias monitoring without manual intervention.

Cost-Effectiveness: AWS Lambda's serverless nature ensures that you only pay for the compute time consumed during the execution of the function, optimizing resource usage.

Flexibility: The solution can be tailored to specific monitoring needs, allowing for adjustments in monitoring frequency and analysis parameters.

By implementing this solution, the company can effectively monitor bias drift in real-time, ensuring that the AI application maintains fairness and accuracy throughout its lifecycle.

Bias drift for models in production - Amazon SageMaker

Schedule Bias Drift Monitoring Jobs - Amazon SageMaker

Question 8

Question Type: MultipleChoice

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?

Options:

A- Low precision
B- High precision
C- Low recall
D- High recall

Answer:

D

Explanation:

Recall measures the ability of a model to correctly identify all positive cases (true positives) out of all actual positives, minimizing false negatives. Since the cost of false negatives is much higher than false positives in this scenario, the ML engineer should prioritize models with high recall to reduce the likelihood of missing positive cases.

Question 9

Question Type: MultipleChoice

A company is using Amazon SageMaker AI to develop a credit risk assessment model. During model validation, the company finds that the model achieves 82% accuracy on the validation data. However, the model achieved 99% accuracy on the training data. The company needs to address the model accuracy issue before deployment.

Which solution will meet this requirement?

Options:

A- Add more dense layers to increase model complexity. Implement batch normalization. Use early stopping during training.
B- Implement dropout layers. Use L1 or L2 regularization. Perform k-fold cross-validation.
C- Use principal component analysis (PCA) to reduce the feature dimensionality. Decrease model layers. Implement cross-entropy loss functions.
D- Augment the training dataset. Remove duplicate records from the training dataset. Implement stratified sampling.

Answer:

B

Explanation:

The large gap between training accuracy (99%) and validation accuracy (82%) is a textbook case of overfitting. The model has learned patterns that fit the training data extremely well but do not generalize to unseen data.

AWS ML best practices recommend regularization techniques to address overfitting. Dropout layers randomly deactivate neurons during training, preventing the network from relying too heavily on specific paths. L1 and L2 regularization penalize large weights, reducing model complexity and improving generalization. k-fold cross-validation provides a more robust

evaluation by training and validating the model across multiple data splits.

Option A increases complexity, which would worsen overfitting. Option C mixes valid ideas (dimensionality reduction) with unrelated changes (loss function choice) and is less targeted. Option D focuses on data quality but does not directly address model variance.

Therefore, implementing dropout, regularization, and k-fold cross-validation is the correct solution.

P2P exams