Exam : **MLA-C01**

Title : AWS Certified Machine Learning Engineer - Associate

Vendor : Amazon

Version : V12.35

NO.1 A company has an ML model that generates text descriptions based on images that customers upload to the company's website. The images can be up to 50 MB in total size.
An ML engineer decides to store the images in an Amazon S3 bucket. The ML engineer must implement a processing solution that can scale to accommodate changes in demand.
Which solution will meet these requirements with the LEAST operational overhead?

A. Create an Amazon SageMaker batch transform job to process all the images in the S3 bucket.

B. Create an Amazon SageMaker Asynchronous Inference endpoint and a scaling policy. Run a script to make an inference request for each image.

C. Create an Amazon Elastic Kubernetes Service (Amazon EKS) cluster that uses Karpenter for auto scaling. Host the model on the EKS cluster. Run a script to make an inference request for each image.

D. Create an AWS Batch job that uses an Amazon Elastic Container Service (Amazon ECS) cluster.Specify a list of images to process for each AWS Batch job.

Answer: B

Explanation:
SageMaker Asynchronous Inference is designed for processing large payloads, such as images up to 50 MB, and can handle requests that do not require an immediate response.
It scales automatically based on the demand, minimizing operational overhead while ensuring cost-efficiency.
A script can be used to send inference requests for each image, and the results can be retrieved asynchronously. This approach is ideal for accommodating varying levels of traffic with minimal manual intervention.

NO.2 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?

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.
References:
* Amazon SageMaker Endpoints
* SageMaker Production Variants
* Monitoring SageMaker Endpoints with CloudWatch

NO.3 A company has a Retrieval Augmented Generation (RAG) application that uses a vector database to store embeddings of documents. The company must migrate the application to AWS and must implement a solution that provides semantic search of text files. The company has already migrated the text repository to an Amazon S3 bucket.
Which solution will meet these requirements?
A. Use an AWS Batch job to process the files and generate embeddings. Use AWS Glue to store the

embeddings. Use SQL queries to perform the semantic searches.

B. Use a custom Amazon SageMaker notebook to run a custom script to generate embeddings. Use SageMaker Feature Store to store the embeddings. Use SQL queries to perform the semantic searches.

C. Use the Amazon Kendra S3 connector to ingest the documents from the S3 bucket into Amazon Kendra. Query Amazon Kendra to perform the semantic searches.

D. Use an Amazon Textract asynchronous job to ingest the documents from the S3 bucket. Query Amazon Textract to perform the semantic searches.

Answer: C

Explanation:

Amazon Kendrais an AI-powered search service designed for semantic search use cases. It allows ingestion of documents from an Amazon S3 bucket using theAmazon Kendra S3 connector. Once the documents are ingested, Kendra enables semantic searches with its built-in capabilities, removing the need to manually generate embeddings or manage a vector database. This approach is efficient, requires minimal operational effort, and meets the requirements for a Retrieval Augmented Generation (RAG) application.

NO.4 An ML engineer is using Amazon SageMaker to train a deep learning model that requires distributed training.

After some training attempts, the ML engineer observes that the instances are not performing as expected. The ML engineer identifies communication overhead between the training instances.

What should the ML engineer do to MINIMIZE the communication overhead between the instances?

A. Place the instances in the same VPC subnet. Store the data in a different AWS Region from where the instances are deployed.

B. Place the instances in the same VPC subnet but in different Availability Zones. Store the data in a different AWS Region from where the instances are deployed.

C. Place the instances in the same VPC subnet. Store the data in the same AWS Region and Availability Zone where the instances are deployed.

D. Place the instances in the same VPC subnet. Store the data in the same AWS Region but in a different Availability Zone from where the instances are deployed.

Answer: C

Explanation:

To minimize communication overhead during distributed training:

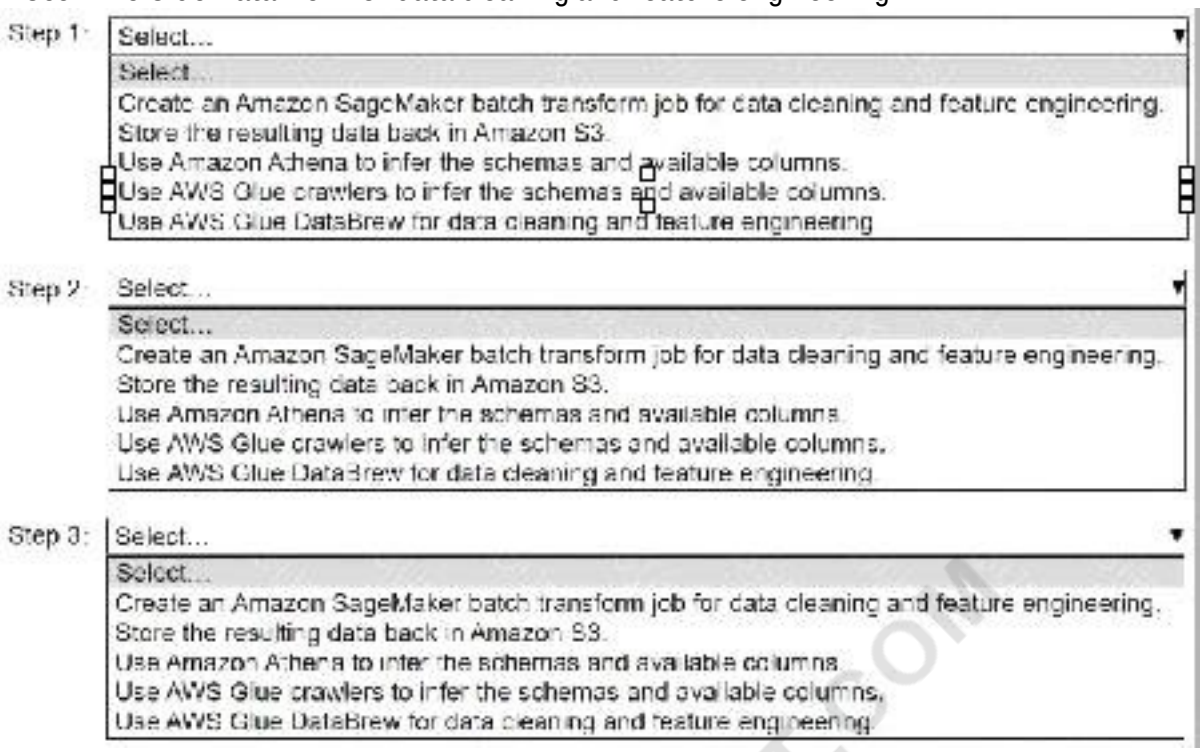
1. Same VPC Subnet: Ensures low-latency communication between training instances by keeping the network traffic within a single subnet.
2. Same AWS Region and Availability Zone: Reduces network latency further because cross-AZ communication incurs additional latency and costs.
3. Data in the Same Region and AZ: Ensures that the training data is accessed with minimal latency, improving performance during training.

This configuration optimizes communication efficiency and minimizes overhead.

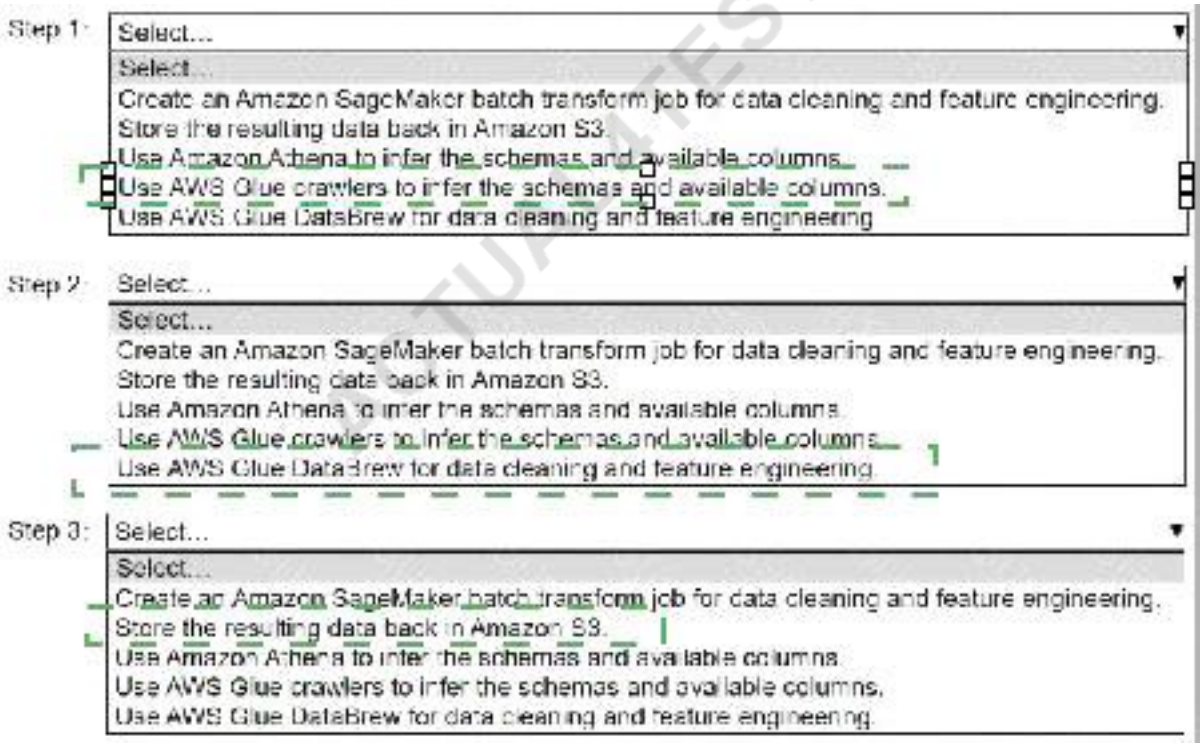
NO.5 A company stores historical data in .csv files in Amazon S3. Only some of the rows and columns in the .csv files are populated. The columns are not labeled. An ML engineer needs to prepare and store the data so that the company can use the data to train ML models.

Select and order the correct steps from the following list to perform this task. Each step should be

selected one time or not at all. (Select and order three.)
* Create an Amazon SageMaker batch transform job for data cleaning and feature engineering.
* Store the resulting data back in Amazon S3.
* Use Amazon Athena to infer the schemas and available columns.
* Use AWS Glue crawlers to infer the schemas and available columns.
* Use AWS Glue DataBrew for data cleaning and feature engineering.



Answer:



Explanation:
Step 1: Use AWS Glue crawlers to infer the schemas and available columns.Step 2: Use AWS Glue DataBrew for data cleaning and feature engineering.Step 3: Store the resulting data back in Amazon

S3.
* Step 1: Use AWS Glue Crawlers to Infer Schemas and Available Columns
* Why?The data is stored in .csv files with unlabeled columns, and Glue Crawlers can scan the raw data in Amazon S3 to automatically infer the schema, including available columns, data types, and any missing or incomplete entries.
* How?Configure AWS Glue Crawlers to point to the S3 bucket containing the .csv files, and run the crawler to extract metadata. The crawler creates a schema in the AWS Glue Data Catalog, which can then be used for subsequent transformations.
* Step 2: Use AWS Glue DataBrew for Data Cleaning and Feature Engineering
* Why?Glue DataBrew is a visual data preparation tool that allows for comprehensive cleaning and transformation of data. It supports imputation of missing values, renaming columns, feature engineering, and more without requiring extensive coding.
* How?Use Glue DataBrew to connect to the inferred schema from Step 1 and perform data cleaning and feature engineering tasks like filling in missing rows/columns, renaming unlabeled columns, and creating derived features.
* Step 3: Store the Resulting Data Back in Amazon S3
* Why?After cleaning and preparing the data, it needs to be saved back to Amazon S3 so that it can be used for training machine learning models.
* How?Configure Glue DataBrew to export the cleaned data to a specific S3 bucket location. This ensures the processed data is readily accessible for ML workflows.
Order Summary:
* Use AWS Glue crawlers to infer schemas and available columns.
* Use AWS Glue DataBrew for data cleaning and feature engineering.
* Store the resulting data back in Amazon S3.
This workflow ensures that the data is prepared efficiently for ML model training while leveraging AWS services for automation and scalability.

NO.6 A company has historical data that shows whether customers needed long-term support from company staff.
The company needs to develop an ML model to predict whether new customers will require long-term support.
Which modeling approach should the company use to meet this requirement?

A. Anomaly detection

B. Linear regression

C. Logistic regression

D. Semantic segmentation

Answer: C
Explanation:
Logistic regression is a suitable modeling approach for this requirement because it is designed for binary classification problems, such as predicting whether a customer will require long-term support ("yes" or "no").
It calculates the probability of a particular class and is widely used for tasks like this where the outcome is categorical.

NO.7 Case study
An ML engineer is developing a fraud detection model on AWS. The training dataset includes

transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
The ML engineer needs to use an Amazon SageMaker built-in algorithm to train the model.
Which algorithm should the ML engineer use to meet this requirement?

A. LightGBM

B. Linear learner

C. #-means clustering

D. Neural Topic Model (NTM)

Answer: B

Explanation:
Why Linear Learner?
* SageMaker'sLinear Learneralgorithm is well-suited for binary classification problems such as fraud detection. It handles class imbalance effectively by incorporating built-in options forweight balancing across classes.
* Linear Learner can capture patterns in the data while being computationally efficient.
Key Features of Linear Learner:
* Automatically weights minority and majority classes.
* Supports both classification and regression tasks.
* Handles interdependencies among features effectively through gradient optimization.
Steps to Implement:
* Use the SageMaker Python SDK to set up a training job with the Linear Learner algorithm.
* Configure the hyperparameters to enable balanced class weights.
* Train the model with the balanced dataset created using SageMaker Data Wrangler.

NO.8 A company has used Amazon SageMaker to deploy a predictive ML model in production. The company is using SageMaker Model Monitor on the model. After a model update, an ML engineer notices data quality issues in the Model Monitor checks.
What should the ML engineer do to mitigate the data quality issues that Model Monitor has identified?

A. Adjust the model's parameters and hyperparameters.

B. Initiate a manual Model Monitor job that uses the most recent production data.

C. Create a new baseline from the latest dataset. Update Model Monitor to use the new baseline for evaluations.

D. Include additional data in the existing training set for the model. Retrain and redeploy the model.

Answer: C

Explanation:
When Model Monitor identifies data quality issues, it might be due to a shift in the data distribution compared to the original baseline. By creating a new baseline using the most recent production data and updating Model Monitor to evaluate against this baseline, the ML engineer ensures that the monitoring is aligned with the current data patterns. This approach mitigates false positives and reflects the updated data characteristics without immediately retraining the model.

NO.9 A company has a conversational AI assistant that sends requests through Amazon Bedrock to an Anthropic Claude large language model (LLM). Users report that when they ask similar questions multiple times, they sometimes receive different answers. An ML engineer needs to improve the responses to be more consistent and less random.
Which solution will meet these requirements?

A. Increase the temperature parameter and the top_k parameter.

B. Increase the temperature parameter. Decrease the top_k parameter.

C. Decrease the temperature parameter. Increase the top_k parameter.

D. Decrease the temperature parameter and the top_k parameter.

Answer: D

Explanation:
Thetemperatureparameter controls the randomness in the model's responses. Lowering the temperature makes the model produce more deterministic and consistent answers.
Thetop_kparameter limits the number of tokens considered for generating the next word. Reducing top_k further constrains the model's options, ensuring more predictable responses.
By decreasing both parameters, the responses become more focused and consistent, reducing variability in similar queries.

NO.10 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

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 falsepositives in this scenario, the ML engineer should prioritize models with high recall to reduce the likelihood of missing positive cases.

NO.11 A company that has hundreds of data scientists is using Amazon SageMaker to create ML models. The models are in model groups in the SageMaker Model Registry.
The data scientists are grouped into three categories: computer vision, natural language processing (NLP), and speech recognition. An ML engineer needs to implement a solution to organize the existing models into these groups to improve model discoverability at scale. The solution must not affect the integrity of the model artifacts and their existing groupings.
Which solution will meet these requirements?

A. Create a custom tag for each of the three categories. Add the tags to the model packages in the SageMaker Model Registry.

B. Create a model group for each category. Move the existing models into these category model groups.

C. Use SageMaker ML Lineage Tracking to automatically identify and tag which model groups should contain the models.

D. Create a Model Registry collection for each of the three categories. Move the existing model groups into the collections.

Answer: A

Explanation:

Using custom tags allows you to organize and categorize models in the SageMaker Model Registry without altering their existing groupings or affecting the integrity of the model artifacts. Tags are a lightweight and scalable way to improve model discoverability at scale, enabling the data scientists to filter and identify models by category (e.g., computer vision, NLP, speech recognition). This approach meets the requirements efficiently without introducing structural changes to the existing model registry setup.

NO.12 An ML engineer needs to use Amazon SageMaker to fine-tune a large language model (LLM) for text summarization. The ML engineer must follow a low-code no-code (LCNC) approach.
Which solution will meet these requirements?

A. Use SageMaker Studio to fine-tune an LLM that is deployed on Amazon EC2 instances.

B. Use SageMaker Autopilot to fine-tune an LLM that is deployed by a custom API endpoint.

C. Use SageMaker Autopilot to fine-tune an LLM that is deployed on Amazon EC2 instances.

D. Use SageMaker Autopilot to fine-tune an LLM that is deployed by SageMaker JumpStart.

Answer: D

Explanation:

SageMaker JumpStart provides access to pre-trained models, including large language models (LLMs), which can be easily deployed and fine-tuned with a low-code/no-code (LCNC) approach. Using SageMaker Autopilot with JumpStart simplifies the fine-tuning process by automating model optimization and reducing the need for extensive coding, making it the ideal solution for this requirement.

NO.13 An ML engineer normalized training data by using min-max normalization in AWS Glue DataBrew. The ML engineer must normalize the production inference data in the same way as the training data before passing the production inference data to the model for predictions.
Which solution will meet this requirement?

A. Apply statistics from a well-known dataset to normalize the production samples.

B. Keep the min-max normalization statistics from the training set. Use these values to normalize the production samples.

C. Calculate a new set of min-max normalization statistics from a batch of production samples. Use these values to normalize all the production samples.

D. Calculate a new set of min-max normalization statistics from each production sample. Use these values to normalize all the production samples.

Answer: B

Explanation:

To ensure consistency between training and inference, themin-max normalization statistics (min and max values)calculated during training must be retained and applied to normalize production inference data. Using the same statistics ensures that the model receives data in the same scale and distribution as it did during training, avoiding discrepancies that could degrade model performance.

Calculating new statistics from production data would lead to inconsistent normalization and affect predictions.

NO.14 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 (CORS) policies to the S3 buckets.

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

Answer: D

Explanation:
By creating IAM policies with specific permissions, you can restrict access to Amazon S3 buckets or objects based on the user's business group. These policies can be attached to IAM users or IAM roles associated with the ML engineers, ensuring that each engineer can only access training data belonging to their group. This approach is secure, scalable, and aligns with AWS best practices for access control.

NO.15 A company is planning to use Amazon Redshift ML in its primary AWS account. The source data is in an Amazon S3 bucket in a secondary account.
An ML engineer needs to set up an ML pipeline in the primary account to access the S3 bucket in the secondary account. The solution must not require public IPv4 addresses.
Which solution will meet these requirements?

A. Provision a Redshift cluster and Amazon SageMaker Studio in a VPC with no public access enabled in the primary account. Create a VPC peering connection between the accounts. Update the VPC route tables to remove the route to 0.0.0.0/0.

B. Provision a Redshift cluster and Amazon SageMaker Studio in a VPC with no public access enabled in the primary account. Create an AWS Direct Connect connection and a transit gateway. Associate the VPCs from both accounts with the transit gateway. Update the VPC route tables to remove the route to
0.0.0.0/0.

C. Provision a Redshift cluster and Amazon SageMaker Studio in a VPC in the primary account. Create an AWS Site-to-Site VPN connection with two encrypted IPsec tunnels between the accounts. Set up interface VPC endpoints for Amazon S3.

D. Provision a Redshift cluster and Amazon SageMaker Studio in a VPC in the primary account. Create an S3 gateway endpoint. Update the S3 bucket policy to allow IAM principals from the primary account.Set up interface VPC endpoints for SageMaker and Amazon Redshift.

Answer: D

Explanation:
S3 Gateway Endpoint: Allows private access to S3 from within a VPC without requiring a public IPv4 address, ensuring that data transfer between the primary and secondary accounts is secure and private.

Bucket Policy Update: The S3 bucket policy in the secondary account must explicitly allow access from the primary account's IAM principals to provide the necessary permissions.
Interface VPC Endpoints: Required for private communication between the VPC and Amazon SageMaker and Amazon Redshift services, ensuring the solution operates without public internet access.
This configuration meets the requirement to avoid public IPv4 addresses and allows secure and private communication between the accounts.

NO.16 A company needs to run a batch data-processing job on Amazon EC2 instances. The job will run during the weekend and will take 90 minutes to finish running. The processing can handle interruptions. The company will run the job every weekend for the next 6 months.
Which EC2 instance purchasing option will meet these requirements MOST cost-effectively?

A. Spot Instances

B. Reserved Instances

C. On-Demand Instances

D. Dedicated Instances

Answer: A

Explanation:
Scenario:The company needs to run a batch job for 90 minutes every weekend over the next 6 months. The processing can handle interruptions, and cost-effectiveness is a priority.
Why Spot Instances?
* Cost-Effective:Spot Instances provide up to 90% savings compared to On-Demand Instances, making them the most cost-effective option for batch processing.
* Interruption Tolerance:Since the processing can tolerate interruptions, Spot Instances are suitable for this workload.
* Batch-Friendly:Spot Instances can be requested for specific durations or automatically re-requested in case of interruptions.
Steps to Implement:
* Create a Spot Instance Request:
* Use the EC2 console or CLI to request Spot Instances with desired instance type and duration.
* Use Auto Scaling:Configure Spot Instances with an Auto Scaling group to handle instance interruptions and ensure job completion.
* Run the Batch Job:Use tools like AWS Batch or custom scripts to manage the processing.
Comparison with Other Options:
* Reserved Instances:Suitable for predictable, continuous workloads, but less cost-effective for a job that runs only once a week.
* On-Demand Instances:More expensive and unnecessary given the tolerance for interruptions.
* Dedicated Instances:Best for isolation and compliance but significantly more costly.
References:
* Amazon EC2 Spot Instances
* Best Practices for Using Spot Instances
* AWS Batch for Spot Instances

NO.17 An ML engineer is developing a fraud detection model by using the Amazon SageMaker XGBoost algorithm.
The model classifies transactions as either fraudulent or legitimate.

During testing, the model excels at identifying fraud in the training dataset. However, the model is inefficient at identifying fraud in new and unseen transactions.
What should the ML engineer do to improve the fraud detection for new transactions?

A. Increase the learning rate.

B. Remove some irrelevant features from the training dataset.

C. Increase the value of the max_depth hyperparameter.

D. Decrease the value of the max_depth hyperparameter.

Answer: D

Explanation:
A high max_depth value in XGBoost can lead to overfitting, where the model learns the training dataset too well but fails to generalize to new and unseen data. By decreasing the max_depth, the model becomes less complex, reducing overfitting and improving its ability to detect fraud in new transactions. This adjustment helps the model focus on general patterns rather than memorizing specific details in the training data.

NO.18 An ML engineer needs to process thousands of existing CSV objects and new CSV objects that are uploaded.
The CSV objects are stored in a central Amazon S3 bucket and have the same number of columns. One of the columns is a transaction date. The ML engineer must query the data based on the transaction date.
Which solution will meet these requirements with the LEAST operational overhead?

A. Use an Amazon Athena CREATE TABLE AS SELECT (CTAS) statement to create a table based on the transaction date from data in the central S3 bucket. Query the objects from the table.

B. Create a new S3 bucket for processed data. Set up S3 replication from the central S3 bucket to the new S3 bucket. Use S3 Object Lambda to query the objects based on transaction date.

C. Create a new S3 bucket for processed data. Use AWS Glue for Apache Spark to create a job to query the CSV objects based on transaction date. Configure the job to store the results in the new S3 bucket.
Query the objects from the new S3 bucket.

D. Create a new S3 bucket for processed data. Use Amazon Data Firehose to transfer the data from the central S3 bucket to the new S3 bucket. Configure Firehose to run an AWS Lambda function to query the data based on transaction date.

Answer: A

Explanation:
Scenario:The ML engineer needs a low-overhead solution to query thousands of existing and new CSV objects stored in Amazon S3 based on a transaction date.
Why Athena?
* Serverless:Amazon Athena is a serverless query service that allows direct querying of data stored in S3 using standard SQL, reducing operational overhead.
* Ease of Use:By using the CTAS statement, the engineer can create a table with optimized partitions based on the transaction date. Partitioning improves query performance and minimizes costs by scanning only relevant data.
* Low Operational Overhead:No need to manage or provision additional infrastructure. Athena integrates seamlessly with S3, and CTAS simplifies table creation and optimization.
Steps to Implement:

* Organize Data in S3:Store CSV files in a bucket in a consistent format and directory structure if possible.
* Configure Athena:Use the AWS Management Console or Athena CLI to set up Athena to point to the S3 bucket.
* Run CTAS Statement:

```
CREATE TABLE processed_data
WITH (
format = 'PARQUET',
external_location = 's3://processed-bucket/',
partitioned_by = ARRAY['transaction_date']
) AS
SELECT *
FROM input_data;
```

This creates a new table with data partitioned by transaction date.
* Query the Data:Use standard SQL queries to fetch data based on the transaction date.
References:
* Amazon Athena CTAS Documentation
* Partitioning Data in Athena

NO.19 Case study
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
Which AWS service or feature can aggregate the data from the various data sources?

A. Amazon EMR Spark jobs

B. Amazon Kinesis Data Streams

C. Amazon DynamoDB

D. AWS Lake Formation

Answer: A
Explanation:
* Problem Description:
* The dataset includes multiple data sources:
* Transaction logs and customer profiles in Amazon S3.
* Tables in an on-premises MySQL database.
* There is aclass imbalancein the dataset andinterdependenciesamong features that need to be addressed.
* The solution requiresdata aggregationfrom diverse sources for centralized processing.
* Why AWS Lake Formation?
* AWS Lake Formationis designed to simplify the process of aggregating, cataloging, and securing data from various sources, including S3, relational databases, and other on-premises systems.
* It integrates with AWS Glue for data ingestion and ETL (Extract, Transform, Load) workflows, making it a robust choice for aggregating data from Amazon S3 and on-premises MySQL databases.
* How It Solves the Problem:

* Data Aggregation: Lake Formation collects data from diverse sources, such as S3 and MySQL, and consolidates it into a centralized data lake.
* Cataloging and Discovery: Automatically crawls and catalogs the data into a searchable catalog, which the ML engineer can query for analysis or modeling.
* Data Transformation: Prepares data using Glue jobs to handle preprocessing tasks such as addressing class imbalance (e.g., oversampling, undersampling) and handling interdependencies among features.
* Security and Governance: Offers fine-grained access control, ensuring secure and compliant data management.
* Steps to Implement Using AWS Lake Formation:
* Step 1: Set up Lake Formation and register data sources, including the S3 bucket and on- premises MySQL database.
* Step 2: Use AWS Glue to create ETL jobs to transform and prepare data for the ML pipeline.
* Step 3: Query and access the consolidated data lake using services such as Athena or SageMaker for further ML processing.
* Why Not Other Options?
* Amazon EMR Spark jobs: While EMR can process large-scale data, it is better suited for complex big data analytics tasks and does not inherently support data aggregation across sources like Lake Formation.
* Amazon Kinesis Data Streams: Kinesis is designed for real-time streaming data, not batch data aggregation across diverse sources.
* Amazon DynamoDB: DynamoDB is a NoSQL database and is not suitable for aggregating data from multiple sources like S3 and MySQL.
Conclusion: AWS Lake Formation is the most suitable service for aggregating data from S3 and on-premises MySQL databases, preparing the data for downstream ML tasks, and addressing challenges like class imbalance and feature interdependencies.
References:
* AWS Lake Formation Documentation
* AWS Glue for Data Preparation

NO.20 A company has an application that uses different APIs to generate embeddings for input text. The company needs to implement a solution to automatically rotate the API tokens every 3 months. Which solution will meet this requirement?

A. Store the tokens in AWS Secrets Manager. Create an AWS Lambda function to perform the rotation.

B. Store the tokens in AWS Systems Manager Parameter Store. Create an AWS Lambda function to perform the rotation.

C. Store the tokens in AWS Key Management Service (AWS KMS). Use an AWS managed key to perform the rotation.

D. Store the tokens in AWS Key Management Service (AWS KMS). Use an AWS owned key to perform the rotation.

Answer: A

Explanation:
AWS Secrets Manager is designed for securely storing, managing, and automatically rotating secrets, including API tokens. By configuring a Lambda function for custom rotation logic, the solution can automatically rotate the API tokens every 3 months as required. Secrets Manager simplifies secret

management and integrates seamlessly with other AWS services, making it the ideal choice for this use case.

NO.21 A credit card company has a fraud detection model in production on an Amazon SageMaker endpoint. The company develops a new version of the model. The company needs to assess the new model's performance by using live data and without affecting production end users.
Which solution will meet these requirements?

A. Set up SageMaker Debugger and create a custom rule.

B. Set up blue/green deployments with all-at-once traffic shifting.

C. Set up blue/green deployments with canary traffic shifting.

D. Set up shadow testing with a shadow variant of the new model.

Answer: D

Explanation:
Shadow testing allows you to send a copy of live production traffic to a shadow variant of the new model while keeping the existing production model unaffected. This enables you to evaluate the performance of the new model in real-time with live data without impacting end users. SageMaker endpoints support this setup by allowing traffic mirroring to the shadow variant, making it an ideal solution for assessing the new model's performance.

NO.22 A company is using ML to predict the presence of a specific weed in a farmer's field. The company is using the Amazon SageMaker linear learner built-in algorithm with a value of multiclass_dassifier for the predictorjype hyperparameter.
What should the company do to MINIMIZE false positives?

A. Set the value of the weight decay hyperparameter to zero.

B. Increase the number of training epochs.

C. Increase the value of the target_precision hyperparameter.

D. Change the value of the predictorjype hyperparameter to regressor.

Answer: C

Explanation:
Thetarget_precisionhyperparameter in the Amazon SageMaker linear learner controls the trade-off between precision and recall for the model. Increasing the target_precision prioritizes minimizing false positives by making the model more cautious in its predictions. This approach is effective for use cases where false positives have higher consequences than false negatives.

NO.23 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/# testing in Amazon CloudWatch.

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

Answer: B

Explanation:

SageMaker Clarify is designed to provide explainability for ML models. It can analyze feature importance and explain how input features influence the model's predictions. By using Clarify with the deployed SageMaker model, the ML engineer can generate insights and present them to stakeholders to explain the sentiment analysis predictions effectively.

NO.24 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?

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.
A white background with black text Description automatically generated

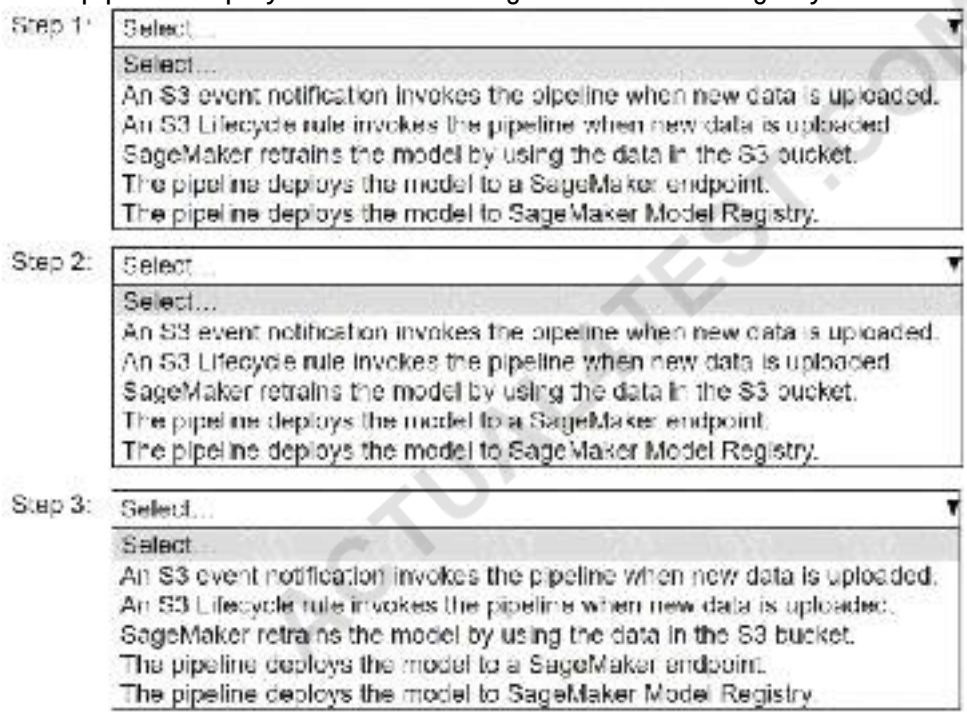
Formula: $\mathrm{MAE} = \frac{1}{n}\sum_{i-1}^{n}|y_i - \hat{y}_i|$ Where:

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

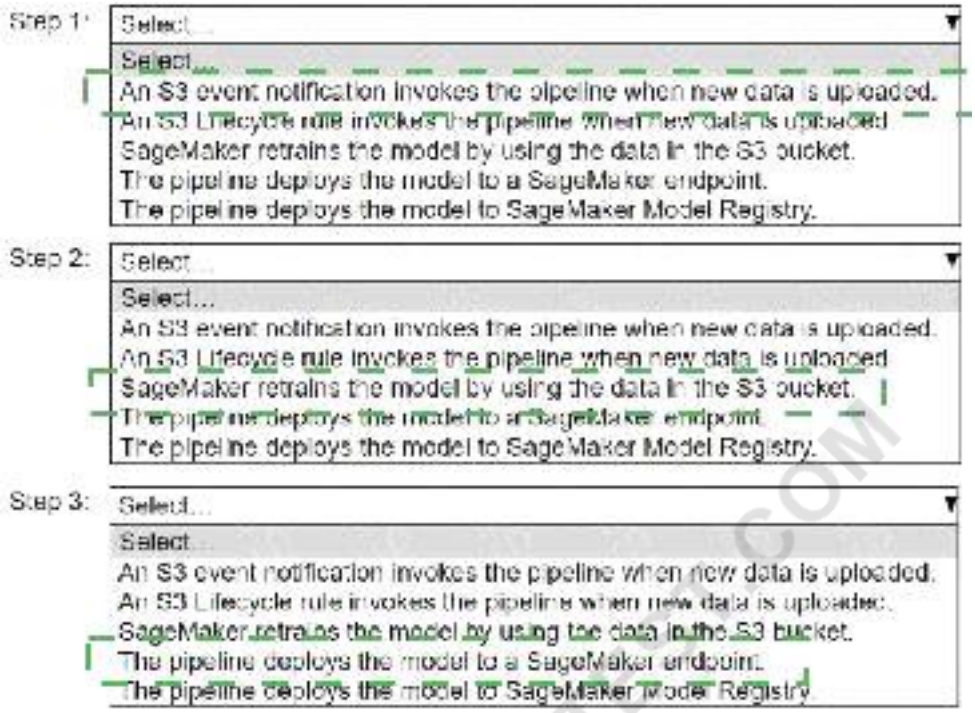
Advantages of MAE:
* Interpretability:MAE is expressed in the same units as the target variable, making it easy to understand.
* Robustness to Outliers:Unlike metrics that square the errors (e.g., Mean Squared Error), MAE does not disproportionately penalize larger errors, making it more robust to outliers.
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.
References:
* Regression Metrics - GeeksforGeeks
* Evaluation Metrics for Your Regression Model - Analytics Vidhya
* Regression Metrics for Machine Learning - Machine Learning Mastery

NO.25 A company wants to host an ML model on Amazon SageMaker. An ML engineer is configuring a continuous integration and continuous delivery (CI/CD) pipeline in AWS CodePipeline to deploy the model. The pipeline must run automatically when new training data for the model is uploaded to an Amazon S3 bucket.
Select and order the pipeline's correct steps from the following list. Each step should be selected one time or not at all. (Select and order three.)
* An S3 event notification invokes the pipeline when new data is uploaded.
* S3 Lifecycle rule invokes the pipeline when new data is uploaded.
* SageMaker retrains the model by using the data in the S3 bucket.
* The pipeline deploys the model to a SageMaker endpoint.
* The pipeline deploys the model to SageMaker Model Registry.

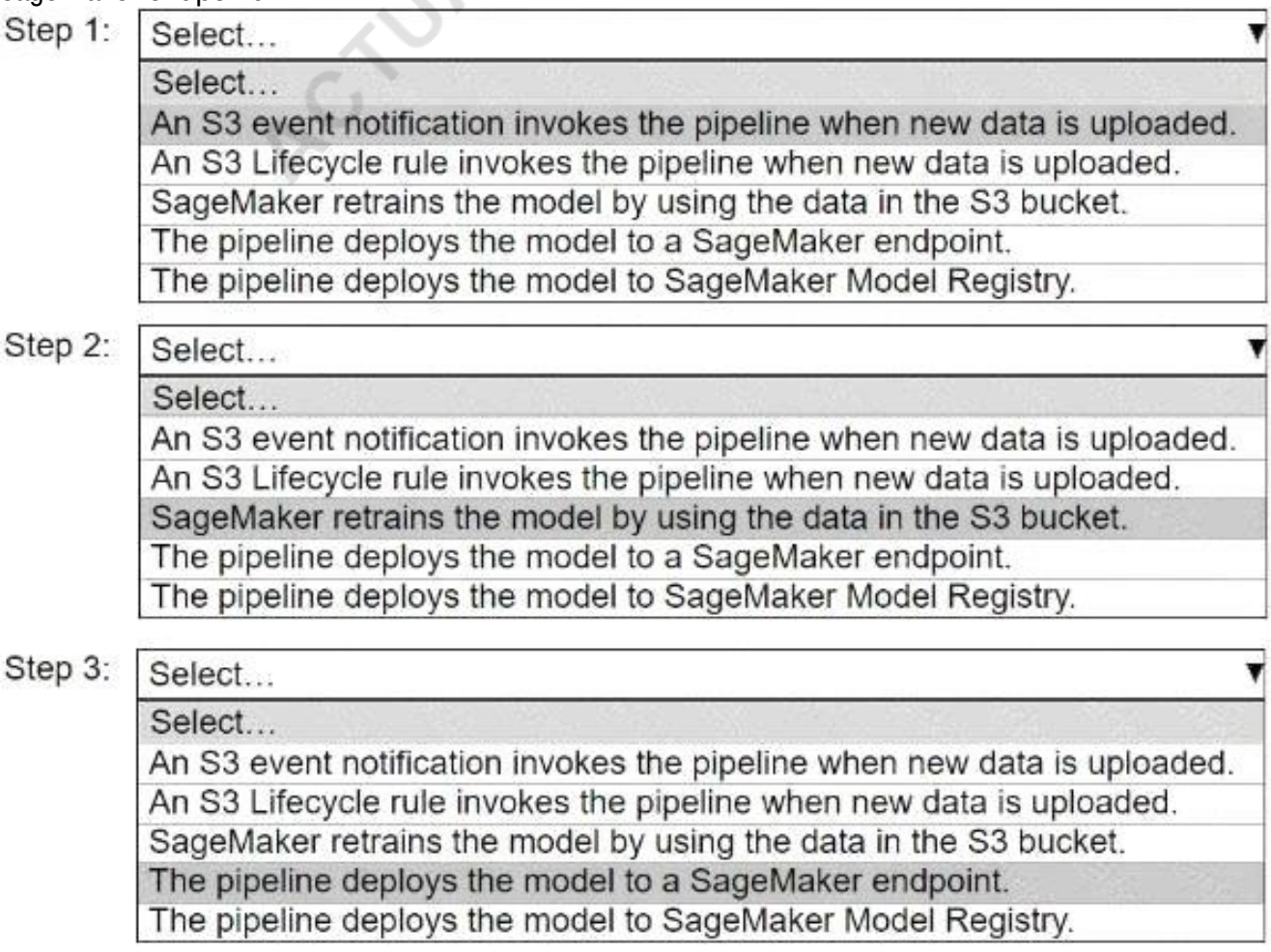


Answer:



Explanation:

Step 1: An S3 event notification invokes the pipeline when new data is uploaded.Step 2: SageMaker retrains the model by using the data in the S3 bucket.Step 3: The pipeline deploys the model to a SageMaker endpoint.



* Step 1: An S3 Event Notification Invokes the Pipeline When New Data is Uploaded

* Why?The CI/CD pipeline should be triggered automatically whenever new training data is uploaded to Amazon S3. S3 event notifications can be configured to send events to AWS services like Lambda, which can then invoke AWS CodePipeline.
* How?Configure the S3 bucket to send event notifications (e.g., s3:ObjectCreated:*) to AWS Lambda, which in turn triggers the CodePipeline.
* Step 2: SageMaker Retrains the Model by Using the Data in the S3 Bucket
* Why?The uploaded data is used to retrain the ML model to incorporate new information and maintain performance. This step is critical to updating the model with fresh data.
* How?Define a SageMaker training step in the CI/CD pipeline, which reads the training data from the S3 bucket and retrains the model.
* Step 3: The Pipeline Deploys the Model to a SageMaker Endpoint
* Why?Once retrained, the updated model must be deployed to a SageMaker endpoint to make it available for real-time inference.
* How?Add a deployment step in the CI/CD pipeline, which automates the creation or update of the SageMaker endpoint with the retrained model.
Order Summary:
* An S3 event notification invokes the pipeline when new data is uploaded.
* SageMaker retrains the model by using the data in the S3 bucket.
* The pipeline deploys the model to a SageMaker endpoint.
This configuration ensures an automated, efficient, and scalable CI/CD pipeline for continuous retraining and deployment of the ML model in Amazon SageMaker.

NO.26 Case study
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
After the data is aggregated, the ML engineer must implement a solution to automatically detect anomalies in the data and to visualize the result.
Which solution will meet these requirements?

A. Use Amazon Athena to automatically detect the anomalies and to visualize the result.

B. Use Amazon Redshift Spectrum to automatically detect the anomalies. Use Amazon QuickSight to visualize the result.

C. Use Amazon SageMaker Data Wrangler to automatically detect the anomalies and to visualize the result.

D. Use AWS Batch to automatically detect the anomalies. Use Amazon QuickSight to visualize the result.

Answer: C
Explanation:
Amazon SageMaker Data Wrangler is a comprehensive tool that streamlines the process of data preparation and offers built-in capabilities for anomaly detection and visualization.
Key Features of SageMaker Data Wrangler:
* Data Importation: Connects seamlessly to various data sources, including Amazon S3 and on-premises databases, facilitating the aggregation of transaction logs, customer profiles, and MySQL

tables.
* Anomaly Detection: Provides built-in analyses to detect anomalies in time series data, enabling the identification of outliers that may indicate fraudulent activities.
* Visualization: Offers a suite of visualization tools, such as histograms and scatter plots, to help understand data distributions and relationships, which are crucial for feature engineering and model development.
Implementation Steps:
* Data Aggregation:
* Import data from Amazon S3 and on-premises MySQL databases into SageMaker Data Wrangler.
* Utilize Data Wrangler's data flow interface to combine and preprocess datasets, ensuring a unified dataset for analysis.
* Anomaly Detection:
* Apply the anomaly detection analysis feature to identify outliers in the dataset.
* Configure parameters such as the anomaly threshold to fine-tune the detection sensitivity.
* Visualization:
* Use built-in visualization tools to create charts and graphs that depict data distributions and highlight anomalies.
* Interpret these visualizations to gain insights into potential fraud patterns and feature interdependencies.
Advantages of Using SageMaker Data Wrangler:
* Integrated Workflow: Combines data preparation, anomaly detection, and visualization within a single interface, streamlining the ML development process.
* Operational Efficiency: Reduces the need for multiple tools and complex integrations, thereby minimizing operational overhead.
* Scalability: Handles large datasets efficiently, making it suitable for extensive transaction logs and customer profiles.
By leveraging SageMaker Data Wrangler, the ML engineer can effectively detect anomalies and visualize results, facilitating the development of a robust fraud detection model.
References:
* Analyze and Visualize - Amazon SageMaker
* Transform Data - Amazon SageMaker

NO.27 A company uses Amazon Athena to query a dataset in Amazon S3. The dataset has a target variable that the company wants to predict.
The company needs to use the dataset in a solution to determine if a model can predict the target variable.
Which solution will provide this information with the LEAST development effort?
A. Create a new model by using Amazon SageMaker Autopilot. Report the model's achieved performance.
B. Implement custom scripts to perform data pre-processing, multiple linear regression, and performance evaluation. Run the scripts on Amazon EC2 instances.
C. Configure Amazon Macie to analyze the dataset and to create a model. Report the model's achieved performance.
D. Select a model from Amazon Bedrock. Tune the model with the data. Report the model's achieved performance.
Answer: A

Explanation:
Amazon SageMaker Autopilot automates the process of building, training, and tuning machine learning models. It provides insights into whether the target variable can be effectively predicted by evaluating the model's performance metrics. This solution requires minimal development effort as SageMaker Autopilot handles data preprocessing, algorithm selection, and hyperparameter optimization automatically, making it the most efficient choice for this scenario.

NO.28 A company uses Amazon SageMaker for its ML workloads. The company's ML engineer receives a 50 MB Apache Parquet data file to build a fraud detection model. The file includes several correlated columns that are not required.
What should the ML engineer do to drop the unnecessary columns in the file with the LEAST effort?
A. Download the file to a local workstation. Perform one-hot encoding by using a custom Python script.
B. Create an Apache Spark job that uses a custom processing script on Amazon EMR.
C. Create a SageMaker processing job by calling the SageMaker Python SDK.
D. Create a data flow in SageMaker Data Wrangler. Configure a transform step.
Answer: D
Explanation:
SageMaker Data Wrangler provides a no-code/low-code interface for preparing and transforming data, including dropping unnecessary columns. By creating a data flow and configuring a transform step, the ML engineer can easily remove correlated or unneeded columns from the Parquet file with minimal effort. This approach avoids the need for custom coding or managing additional infrastructure.

NO.29 A company has a team of data scientists who use Amazon SageMaker notebook instances to test ML models.
When the data scientists need new permissions, the company attaches the permissions to each individual role that was created during the creation of the SageMaker notebook instance.
The company needs to centralize management of the team's permissions.
Which solution will meet this requirement?
A. Create a single IAM role that has the necessary permissions. Attach the role to each notebook instance that the team uses.
B. Create a single IAM group. Add the data scientists to the group. Associate the group with each notebook instance that the team uses.
C. Create a single IAM user. Attach the AdministratorAccess AWS managed IAM policy to the user. Configure each notebook instance to use the IAM user.
D. Create a single IAM group. Add the data scientists to the group. Create an IAM role. Attach the AdministratorAccess AWS managed IAM policy to the role. Associate the role with the group.Associate the group with each notebook instance that the team uses.
Answer: A
Explanation:
Managing permissions for multiple Amazon SageMaker notebook instances can become complex when handled individually. To centralize and streamline permission management, AWS recommends creating a single IAM role with the necessary permissions and attaching this role to each notebook instance used by the data science team.

Steps to Implement the Solution:
* Create a Single IAM Role with Necessary Permissions:
* Define an IAM role that encompasses all permissions required by the data scientists for their tasks. This includes permissions for SageMaker operations and any other AWS services they interact with.
* AWS provides managed policies like AmazonSageMakerFullAccess that can be attached to the role to grant comprehensive SageMaker permissions.(IAM Policies for SageMaker)
* Attach the IAM Role to Each Notebook Instance:
* When creating or updating a SageMaker notebook instance, specify the IAM role created in the previous step. This ensures that all notebook instances operate under a consistent set of permissions.
* In the SageMaker console, during the notebook instance setup, you can choose an existing IAM role to associate with the instance.(Creating SageMaker Workspaces) Benefits of This Approach:
* Centralized Permission Management:By using a single IAM role, you simplify the process of updating permissions. Changes to the role's policies automatically propagate to all associated notebook instances, ensuring consistent access control.
* Adherence to Best Practices:AWS recommends using IAM roles to manage permissions for applications running on services like SageMaker. This approach avoids the need to manage individual user permissions separately.(IAM Best Practices for SageMaker) Alternative Options and Their Drawbacks:
* Option B:Creating a single IAM group and adding data scientists to it does not directly associate the group with notebook instances. IAM groups are used to manage user permissions, not to assign roles to AWS resources like notebook instances.
* Option C:Using a single IAM user with the AdministratorAccess policy is not recommended due to security risks associated with granting broad permissions and the challenges in managing shared user credentials.
* Option D:Associating an IAM group with a role and then with notebook instances is not a valid approach, as IAM groups cannot be directly associated with AWS resources.
Conclusion:Option A is the most effective solution to centralize and manage permissions for SageMaker notebook instances, aligning with AWS best practices for IAM role management.
References:
* AWS Documentation: IAM Policies for SageMaker
* AWS Documentation: Creating SageMaker Workspaces
* AWS Documentation: IAM Best Practices for SageMaker

NO.30 An ML engineer has trained a neural network by using stochastic gradient descent (SGD). The neural network performs poorly on the test set. The values for training loss and validation loss remain high and show an oscillating pattern. The values decrease for a few epochs and then increase for a few epochs before repeating the same cycle.
What should the ML engineer do to improve the training process?

A. Introduce early stopping.

B. Increase the size of the test set.

C. Increase the learning rate.

D. Decrease the learning rate.

Answer: D
Explanation:
In training neural networks using Stochastic Gradient Descent (SGD), the learning rate is a critical hyperparameter that influences the convergence behavior of the model. Observing oscillations in

training and validation loss suggests that the learning rate may be too high, causing the optimization process to overshoot minima in the loss landscape.
Understanding the Impact of Learning Rate:
* High Learning Rate:A high learning rate can cause the model parameters to update too aggressively, leading to oscillations or divergence in the loss function. This manifests as the loss decreasing for a few epochs and then increasing, repeating this cycle without stable convergence.
* Low Learning Rate:A low learning rate results in smaller parameter updates, allowing the model to converge more steadily to a minimum, albeit potentially at a slower pace.
Recommended Action:
Decreasing the learning rate allows for more precise adjustments to the model parameters, facilitating smoother convergence and reducing oscillations in the loss function. This adjustment helps the model settle into minima more effectively, improving overall performance.
Supporting Evidence:
Research indicates that large learning rates can lead to phenomena such as "catapults," where spikes in training loss occur due to aggressive updates. Reducing the learning rate mitigates these issues, promoting stable training dynamics.
References:
* Catapults in SGD: Spikes in the Training Loss and Their Impact on Generalization Through Feature Learning
* Lecture 7: Training Neural Networks, Part 2 - Stanford University
Conclusion:
To address oscillating training and validation loss during neural network training with SGD, decreasing the learning rate is an effective strategy. This adjustment facilitates smoother convergence and enhances the model's performance on the test set.

NO.31 An ML engineer has an Amazon Comprehend custom model in Account A in the us-east-1 Region. The ML engineer needs to copy the model to Account # in the same Region.
Which solution will meet this requirement with the LEAST development effort?
A. Use Amazon S3 to make a copy of the model. Transfer the copy to Account B.
B. Create a resource-based IAM policy. Use the Amazon Comprehend ImportModel API operation to copy the model to Account B.
C. Use AWS DataSync to replicate the model from Account A to Account B.
D. Create an AWS Site-to-Site VPN connection between Account A and Account # to transfer the model.
Answer: B
Explanation:
Amazon Comprehend provides the ImportModel API operation, which allows you to copy a custom model between AWS accounts. By creating a resource-based IAM policy on the model in Account A, you can grant Account B the necessary permissions to access and import the model. This approach requires minimal development effort and is the AWS-recommended method for sharing custom models across accounts.

NO.32 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?

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.
References:
* Bias drift for models in production - Amazon SageMaker
* Schedule Bias Drift Monitoring Jobs - Amazon SageMaker

NO.33 A company wants to develop an ML model by using tabular data from its customers. The data contains meaningful ordered features with sensitive information that should not be discarded. An ML engineer must ensure that the sensitive data is masked before another team starts to build the

model.
Which solution will meet these requirements?
A. Use Amazon Made to categorize the sensitive data.
B. Prepare the data by using AWS Glue DataBrew.
C. Run an AWS Batch job to change the sensitive data to random values.
D. Run an Amazon EMR job to change the sensitive data to random values.
Answer: B
Explanation:
AWS Glue DataBrew provides an easy-to-use interface for preparing and transforming data, including masking or obfuscating sensitive information. It offers built-in data masking features, allowing the ML engineer to handle sensitive data securely while retaining its structure and meaning. This solution is efficient and requires minimal coding, making it ideal for ensuring sensitive data is masked before model building begins.

NO.34 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
Answer: A
Explanation:
The AWS::SageMaker::Model resource in AWS CloudFormation is used to create an ML model in Amazon SageMaker. This model can then be hosted on an endpoint by using the AWS::SageMaker::Endpoint resource. The model resource defines the container or algorithm to use for hosting and the S3 location of the model artifacts.

NO.35 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
Answer: A
Explanation:
Accuracy measures the proportion of correctly predicted labels (both positive and negative) out of the total predictions. It is the appropriate metric when the goal is to maximize the correct predictions of both positive and negative labels. However, it assumes that the classes are balanced; if the classes are imbalanced, other metrics like precision, recall, or specificity may be more relevant depending on

the specific needs.

NO.36 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 is experimenting with consecutive training jobs.
How can the company MINIMIZE infrastructure startup times for these jobs?

A. Use Managed Spot Training.

B. Use SageMaker managed warm pools.

C. Use SageMaker Training Compiler.

D. Use the SageMaker distributed data parallelism (SMDDP) library.

Answer: B
Explanation:
When running consecutive training jobs in Amazon SageMaker, infrastructure provisioning can introduce latency, as each job typically requires the allocation and setup of compute resources. To minimize this startup time and enhance efficiency, Amazon SageMaker offersManaged Warm Pools.
Key Features of Managed Warm Pools:
* Reduced Latency: Reusing existing infrastructure significantly reduces startup time for training jobs.
* Configurable Retention Period: Allows retention of resources after training jobs complete, defined by the KeepAlivePeriodInSeconds parameter.
* Automatic Matching: Subsequent jobs with matching configurations (e.g., instance type) can reuse retained infrastructure.
Implementation Steps:
* Request Warm Pool Quota Increase: Increase the default resource quota for warm pools through AWS Service Quotas.
* Configure Training Jobs:
* Set KeepAlivePeriodInSeconds for the first training job to retain resources.
* Ensure subsequent jobs match the retained pool's configuration to enable reuse.
* Monitor Warm Pool Usage: Track warm pool status through the SageMaker console or API to confirm resource reuse.
Considerations:
* Billing: Resources in warm pools are billable during the retention period.
* Matching Requirements: Jobs must have consistent configurations to use warm pools effectively.
Alternative Options:
* Managed Spot Training: Reduces costs by using spare capacity but doesn't address startup latency.
* SageMaker Training Compiler: Optimizes training time but not infrastructure setup.
* SageMaker Distributed Data Parallelism Library: Enhances training efficiency but doesn't reduce setup time.
By usingManaged Warm Pools, the company can significantly reduce startup latency for consecutive training jobs, ensuring faster experimentation cycles with minimal operational overhead.
References:
* AWS Documentation: Managed Warm Pools
* AWS Blog: Reduce ML Model Training Job Startup Time

NO.37 A company is creating an application that will recommend products for customers to purchase. The application will make API calls to Amazon Q Business. The company must ensure that responses from Amazon Q Business do not include the name of the company's main competitor. Which solution will meet this requirement?

A. Configure the competitor's name as a blocked phrase in Amazon Q Business.

B. Configure an Amazon Q Business retriever to exclude the competitor's name.

C. Configure an Amazon Kendra retriever for Amazon Q Business to build indexes that exclude the competitor's name.

D. Configure document attribute boosting in Amazon Q Business to deprioritize the competitor's name.

Answer: A

Explanation:

Amazon Q Business allows configuring blocked phrases to exclude specific terms or phrases from the responses. By adding the competitor's name as a blocked phrase, the company can ensure that it will not appear in the API responses, meeting the requirement efficiently with minimal configuration.

NO.38 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 Studio to manage the entire ML workflows. Use SageMaker ML Lineage Tracking 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 Studio to manage the entire ML workflows. Use SageMaker ML Lineage Tracking 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.

Answer: C

Explanation:

SageMaker Pipelines provides a directed acyclic graph (DAG) view for managing and visualizing ML workflows with fine-grained control. It integrates seamlessly with SageMaker Studio, offering an intuitive interface for workflow orchestration.

SageMaker ML Lineage Tracking keeps a running history of experiments and tracks the lineage of datasets, models, and training jobs. This feature supports model governance, auditing, and compliance verification requirements.

NO.39 A company has implemented a data ingestion pipeline for sales transactions from its

ecommerce website. The company uses Amazon Data Firehose to ingest data into Amazon OpenSearch Service. The buffer interval of the Firehose stream is set for 60 seconds. An OpenSearch linear model generates real-time sales forecasts based on the data and presents the data in an OpenSearch dashboard.
The company needs to optimize the data ingestion pipeline to support sub-second latency for the real-time dashboard.
Which change to the architecture will meet these requirements?

A. Use zero buffering in the Firehose stream. Tune the batch size that is used in the PutRecordBatch operation.

B. Replace the Firehose stream with an AWS DataSync task. Configure the task with enhanced fan-out consumers.

C. Increase the buffer interval of the Firehose stream from 60 seconds to 120 seconds.

D. Replace the Firehose stream with an Amazon Simple Queue Service (Amazon SQS) queue.

Answer: A

Explanation:
Amazon Kinesis Data Firehose allows for near real-time data streaming. Setting thebuffering hintsto zero or a very small value minimizes the buffering delay and ensures that records are delivered to the destination (Amazon OpenSearch Service) as quickly as possible. Additionally, tuning thebatch sizein thePutRecordBatchoperation can further optimize the data ingestion for sub-second latency. This approach minimizes latency while maintaining the operational simplicity of using Firehose.

NO.40 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.

Answer: A

Explanation:
Dynamic data maskingallows you to control how sensitive data is presented to users at query time, without modifying or storing transformed versions of the source data. Amazon Redshift supports dynamic data masking, which can be implemented with minimal effort. This solution ensures that the data scientistcan access the required information while sensitive data remains protected, meeting the requirements efficiently and with the least implementation effort.

NO.41 An ML engineer trained an ML model on Amazon SageMaker to detect automobile accidents from dosed- circuit TV footage. The ML engineer used SageMaker Data Wrangler to create a training

dataset of images of accidents and non-accidents.
The model performed well during training and validation. However, the model is underperforming in production because of variations in the quality of the images from various cameras.
Which solution will improve the model's accuracy in the LEAST amount of time?

A. Collect more images from all the cameras. Use Data Wrangler to prepare a new training dataset.

B. Recreate the training dataset by using the Data Wrangler corrupt image transform. Specify the impulse noise option.

C. Recreate the training dataset by using the Data Wrangler enhance image contrast transform. Specify the Gamma contrast option.

D. Recreate the training dataset by using the Data Wrangler resize image transform. Crop all images to the same size.

Answer: B

Explanation:
The model is underperforming in production due to variations in image quality from different cameras. Using the corrupt image transform with the impulse noise option in SageMaker Data Wrangler simulates real-world noise and variations in the training dataset. This approach helps the model become more robust to inconsistencies in image quality, improving its accuracy in production without the need to collect and process new data, thereby saving time.

NO.42 Case study
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
Before the ML engineer trains the model, the ML engineer must resolve the issue of the imbalanced data.
Which solution will meet this requirement with the LEAST operational effort?

A. Use Amazon Athena to identify patterns that contribute to the imbalance. Adjust the dataset accordingly.

B. Use Amazon SageMaker Studio Classic built-in algorithms to process the imbalanced dataset.

C. Use AWS Glue DataBrew built-in features to oversample the minority class.

D. Use the Amazon SageMaker Data Wrangler balance data operation to oversample the minority class.

Answer: D

Explanation:
Problem Description:
* The training dataset has a class imbalance, meaning one class (e.g., fraudulent transactions) has fewer samples compared to the majority class (e.g., non-fraudulent transactions). This imbalance affects the model's ability to learn patterns from the minority class.
Why SageMaker Data Wrangler?
* SageMaker Data Wrangler provides a built-in operation called "Balance Data," which includes oversampling and undersampling techniques to address class imbalances.
* Oversampling the minority class replicates samples of the minority class, ensuring the algorithm

receives balanced inputs without significant additional operational overhead.
Steps to Implement:
* Import the dataset into SageMaker Data Wrangler.
* Apply the "Balance Data" operation and configure it to oversample the minority class.
* Export the balanced dataset for training.
Advantages:
* Ease of Use: Minimal configuration is required.
* Integrated Workflow: Works seamlessly with the SageMaker ecosystem for preprocessing and model training.
* Time Efficiency: Reduces manual effort compared to external tools or scripts.

NO.43 A company is planning to use Amazon SageMaker to make classification ratings that are based on images.
The company has 6 ## of training data that is stored on an Amazon FSx for NetApp ONTAP system virtual machine (SVM). The SVM is in the same VPC as SageMaker.
An ML engineer must make the training data accessible for ML models that are in the SageMaker environment.
Which solution will meet these requirements?
A. Mount the FSx for ONTAP file system as a volume to the SageMaker Instance.
B. Create an Amazon S3 bucket. Use Mountpoint for Amazon S3 to link the S3 bucket to the FSx for ONTAP file system.
C. Create a catalog connection from SageMaker Data Wrangler to the FSx for ONTAP file system.
D. Create a direct connection from SageMaker Data Wrangler to the FSx for ONTAP file system.
Answer: A
Explanation:
Amazon FSx for NetApp ONTAP allows mounting the file system as a network-attached storage (NAS) volume. Since the FSx for ONTAP file system and SageMaker instance are in the same VPC, you can directly mount the file system to the SageMaker instance. This approach ensures efficient access to the 6 TB of training data without the need to duplicate or transfer the data, meeting the requirements with minimal complexity and operational overhead.

NO.44 A company has deployed an XGBoost prediction model in production to predict if a customer is likely to cancel a subscription. The company uses Amazon SageMaker Model Monitor to detect deviations in the F1 score.
During a baseline analysis of model quality, the company recorded a threshold for the F1 score. After several months of no change, the model's F1 score decreases significantly.
What could be the reason for the reduced F1 score?
A. Concept drift occurred in the underlying customer data that was used for predictions.
B. The model was not sufficiently complex to capture all the patterns in the original baseline data.
C. The original baseline data had a data quality issue of missing values.
D. Incorrect ground truth labels were provided to Model Monitor during the calculation of the baseline.
Answer: A
Explanation:
* Problem Description:

* The F1 score, which is a balance of precision and recall, has decreased significantly. This indicates the model's predictions are no longer aligned with the real-world data distribution.
* Why Concept Drift?
* Concept driftoccurs when the statistical properties of the target variable or features change over time. For example, customer behaviors or subscription cancellation patterns may have shifted, leading to reduced model accuracy.
* Signs of Concept Drift:
* Deviation in performance metrics (e.g., F1 score) over time.
* Declining prediction accuracy for certain groups or scenarios.
* Solution:
* Monitor for drift using tools like SageMaker Model Monitor.
* Regularly retrain the model with updated data to account for the drift.
* Why Not Other Options?:
* B: Model complexity is unrelated if the model initially performed well.
* C: Data quality issues would have been detected during baseline analysis.
* D: Incorrect ground truth labels would have resulted in a consistently poor baseline.
Conclusion: The decrease in F1 score is most likely due toconcept driftin the customer data, requiring retraining of the model with new data.

NO.45 A company runs an Amazon SageMaker domain in a public subnet of a newly created VPC. The network is configured properly, and ML engineers can access the SageMaker domain.
Recently, the company discovered suspicious traffic to the domain from a specific IP address. The company needs to block traffic from the specific IP address.
Which update to the network configuration will meet this requirement?

A. Create a security group inbound rule to deny traffic from the specific IP address. Assign the security group to the domain.

B. Create a network ACL inbound rule to deny traffic from the specific IP address. Assign the rule to the default network Ad for the subnet where the domain is located.

C. Create a shadow variant for the domain. Configure SageMaker Inference Recommender to send traffic from the specific IP address to the shadow endpoint.

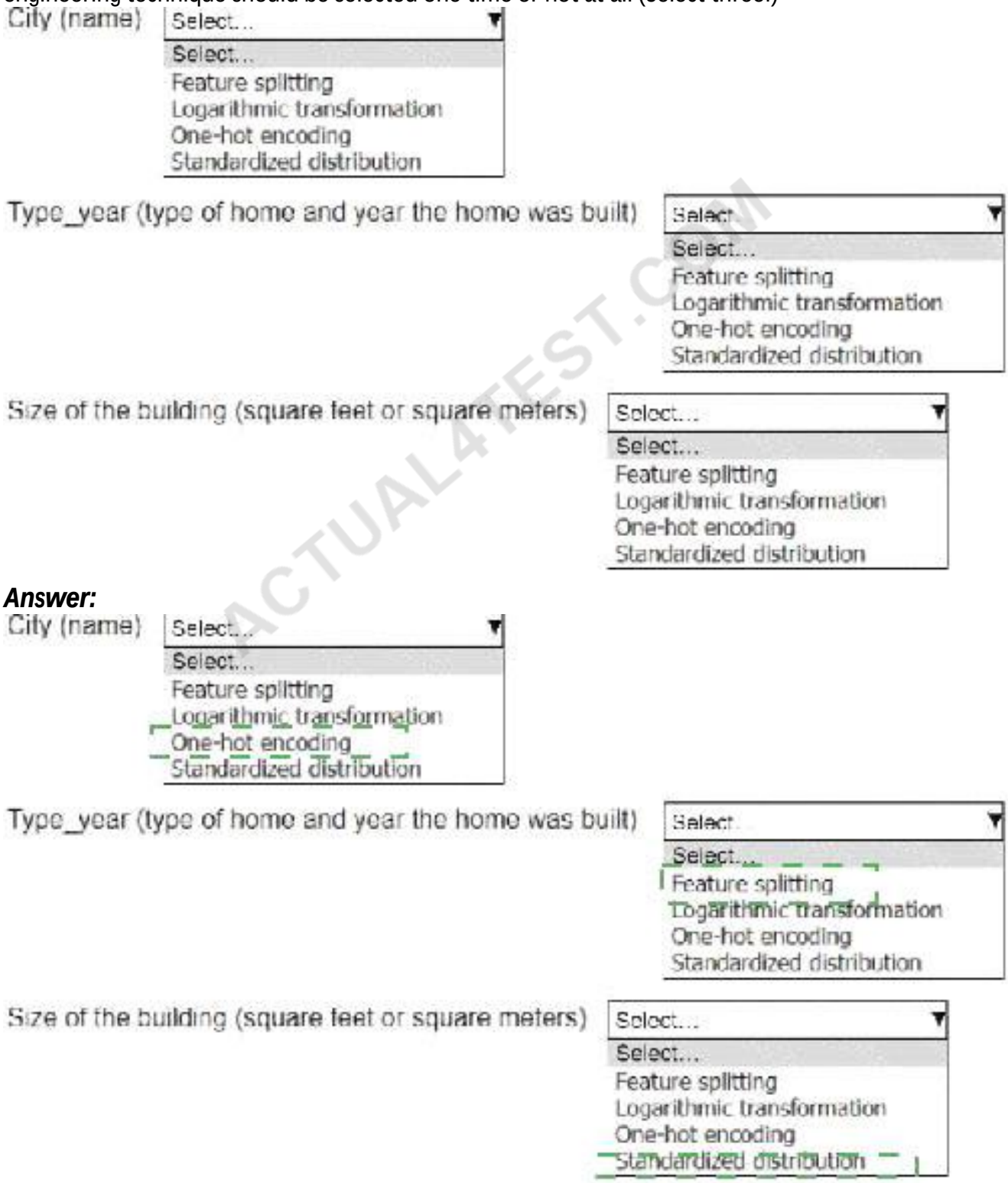
D. Create a VPC route table to deny inbound traffic from the specific IP address. Assign the route table to the domain.

Answer: B
Explanation:
Network ACLs (Access Control Lists) operate at the subnet level and allow for rules to explicitly deny traffic from specific IP addresses. By creating an inbound rule in the network ACL to deny traffic from the suspicious IP address, the company can block traffic to the Amazon SageMaker domain from that IP. This approach works because network ACLs are evaluated before traffic reaches the security groups, making them effective for blocking traffic at the subnet level.

NO.46 An ML engineer is working on an ML model to predict the prices of similarly sized homes. The model will base predictions on several features The ML engineer will use the following feature engineering techniques to estimate the prices of the homes:
* Feature splitting
* Logarithmic transformation
* One-hot encoding

* Standardized distribution

Select the correct feature engineering techniques for the following list of features. Each feature engineering technique should be selected one time or not at all (Select three.)



Answer:

Explanation:

* City (name):One-hot encoding
* Type_year (type of home and year the home was built):Feature splitting
* Size of the building (square feet or square meters):Standardized distribution
* City (name): One-hot encoding

* Why?The "City" is a categorical feature (non-numeric), so one-hot encoding is used to transform it into a numeric format. This encoding creates binary columns for eachunique category (e.g., cities like "New York" or "Los Angeles"), which the model can interpret.
* Type_year (type of home and year the home was built): Feature splitting
* Why?"Type_year" combines two pieces of information into one column, which could confuse the model. Feature splitting separates this column into two distinct features: "Type of home" and "Year built," enabling the model to process each feature independently.
* Size of the building (square feet or square meters): Standardized distribution
* Why?Size is a continuous numerical variable, and standardization (scaling the feature to have a mean of 0 and a standard deviation of 1) ensures that the model treats it fairly compared to other features, avoiding bias from differences in feature scale.
By applying these feature engineering techniques, the ML engineer can ensure that the input data is correctly formatted and optimized for the model to make accurate predictions.

NO.47 A company has a large, unstructured dataset. The dataset includes many duplicate records across several key attributes.
Which solution on AWS will detect duplicates in the dataset with the LEAST code development?
A. Use Amazon Mechanical Turk jobs to detect duplicates.
B. Use Amazon QuickSight ML Insights to build a custom deduplication model.
C. Use Amazon SageMaker Data Wrangler to pre-process and detect duplicates.
D. Use the AWS Glue FindMatches transform to detect duplicates.
Answer: D
Explanation:
Scenario:The dataset contains duplicate records that need to be detected with minimal code development.
Why FindMatches in AWS Glue?
* Purpose-Built for Deduplication:The FindMatches transform in AWS Glue is specifically designed to identify duplicate records in structured or semi-structured datasets.
* Machine Learning-Based:It uses ML to identify duplicates based on configurable thresholds and provides flexibility for tuning accuracy.
* Low Code Overhead:Minimal development effort is required as Glue provides an interactive console for configuring and running FindMatches transforms.
Steps to Implement:
* Prepare the Data:Upload the unstructured dataset to an S3 bucket and define a schema if needed.
* Create a Glue Job:
* Use the AWS Glue Studio to create a job and select the FindMatches transform.
* Specify key attributes for deduplication.
* Run and Evaluate:Execute the Glue job, and review the results for duplicates.
* Resolve Duplicates:Export results to an S3 bucket or process them as needed.
References:
* AWS Glue FindMatches Documentation
* FindMatches Transform Example

NO.48 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 must implement a manual approval-based workflow to ensure that only approved models can be deployed to production endpoints.
Which solution will meet this requirement?

A. Use SageMaker Experiments to facilitate the approval process during model registration.

B. Use SageMaker ML Lineage Tracking on the central model registry. Create tracking entities for the approval process.

C. Use SageMaker Model Monitor to evaluate the performance of the model and to manage the approval.

D. Use SageMaker Pipelines. When a model version is registered, use the AWS SDK to change the approval status to "Approved."

Answer: D

Explanation:
To implement a manual approval-based workflow ensuring that only approved models are deployed to production endpoints, Amazon SageMaker provides integrated tools such asSageMaker Pipelinesand the SageMaker Model Registry.
SageMaker Pipelinesis a robust service for building, automating, and managing end-to-end machine learning workflows. It facilitates the orchestration of various steps in the ML lifecycle, including data preprocessing, model training, evaluation, and deployment. By integrating with theSageMaker Model Registry, it enables seamless tracking and management of model versions and their approval statuses.
Implementation Steps:
* Define the Pipeline:
* Create a SageMaker Pipeline encompassing steps for data preprocessing, model training, evaluation, and registration of the model in the Model Registry.
* Incorporate aCondition Stepto assess model performance metrics. If the model meets predefined criteria, proceed to the next step; otherwise, halt the process.
* Register the Model:
* Utilize theRegisterModelstep to add the trained model to the Model Registry.
* Set the ModelApprovalStatus parameter to PendingManualApproval during registration. This status indicates that the model awaits manual review before deployment.
* Manual Approval Process:
* Notify the designated approver upon model registration. This can be achieved by integrating Amazon EventBridge to monitor registration events and trigger notifications via AWS Lambda functions.
* The approver reviews the model's performance and, if satisfactory, updates the model's status to Approved using the AWS SDK or through the SageMaker Studio interface.
* Deploy the Approved Model:
* Configure the pipeline to automatically deploy models with an Approved status to the production endpoint. This can be managed by adding deployment steps conditioned on the model's approval status.
Advantages of This Approach:
* Automated Workflow:SageMaker Pipelines streamline the ML workflow, reducing manual interventions and potential errors.

* Governance and Compliance:The manual approval step ensures that only thoroughly evaluated models are deployed, aligning with organizational standards.
* Scalability:The solution supports complex ML workflows, making it adaptable to various project requirements.
By implementing this solution, the company can establish a controlled and efficient process for deploying models, ensuring that only approved versions reach production environments.
References:
* Automate the machine learning model approval process with Amazon SageMaker Model Registry and Amazon SageMaker Pipelines
* Update the Approval Status of a Model - Amazon SageMaker

NO.49 An ML engineer needs to create data ingestion pipelines and ML model deployment pipelines on AWS. All the raw data is stored in Amazon S3 buckets.
Which solution will meet these requirements?

A. Use Amazon Data Firehose to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.

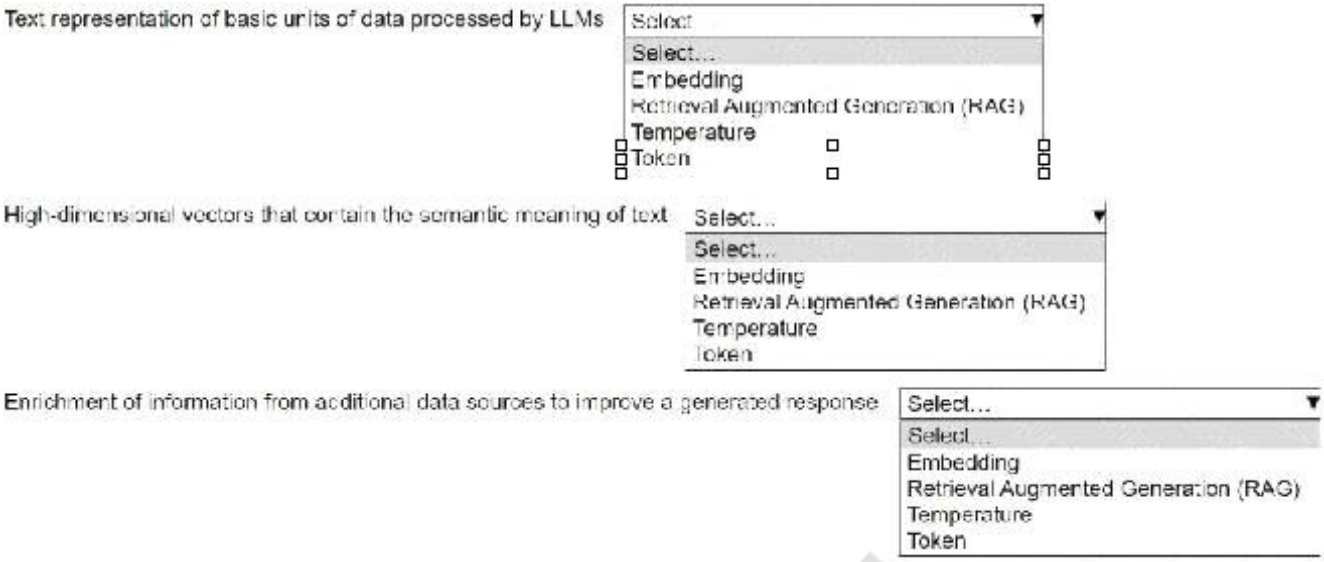
B. Use AWS Glue to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.

C. Use Amazon Redshift ML to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.

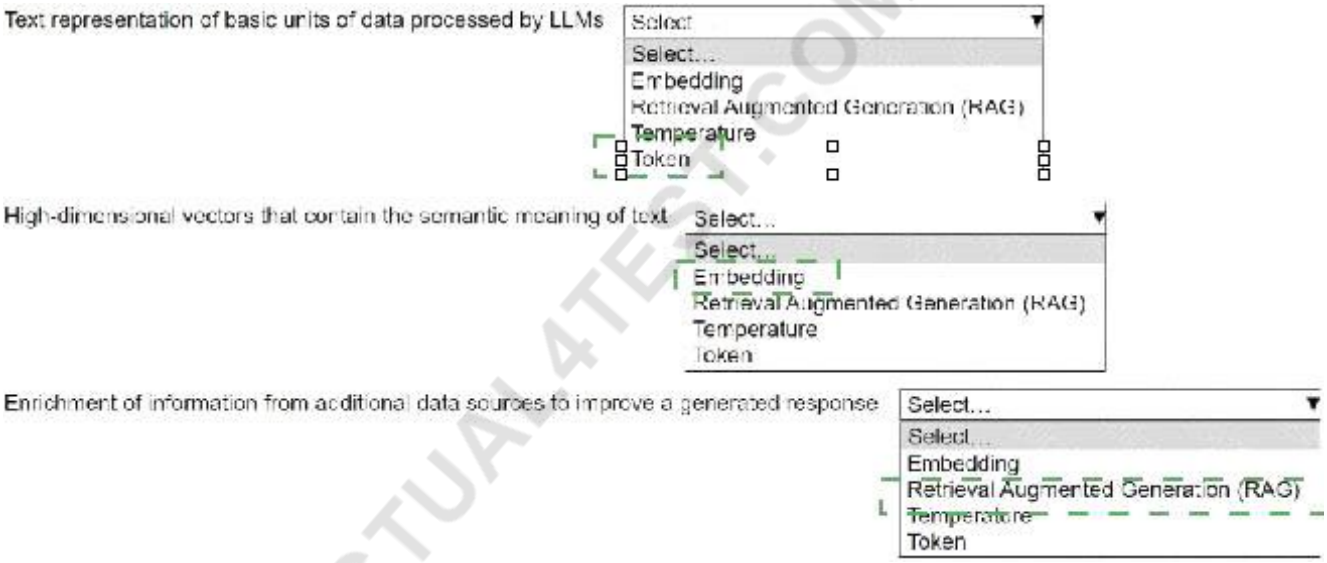
D. Use Amazon Athena to create the data ingestion pipelines. Use an Amazon SageMaker notebook to create the model deployment pipelines.

Answer: B
Explanation:
AWS Glue is a serverless data integration service that is well-suited for creating data ingestion pipelines, especially when raw data is stored in Amazon S3. It can clean, transform, and catalog data, making it accessible for downstream ML tasks.
Amazon SageMaker Studio Classic provides a comprehensive environment for building, training, and deploying ML models. It includes built-in tools and capabilities to create efficient model deployment pipelines with minimal setup.
This combination ensures seamless integration of data ingestion and ML model deployment with minimal operational overhead.

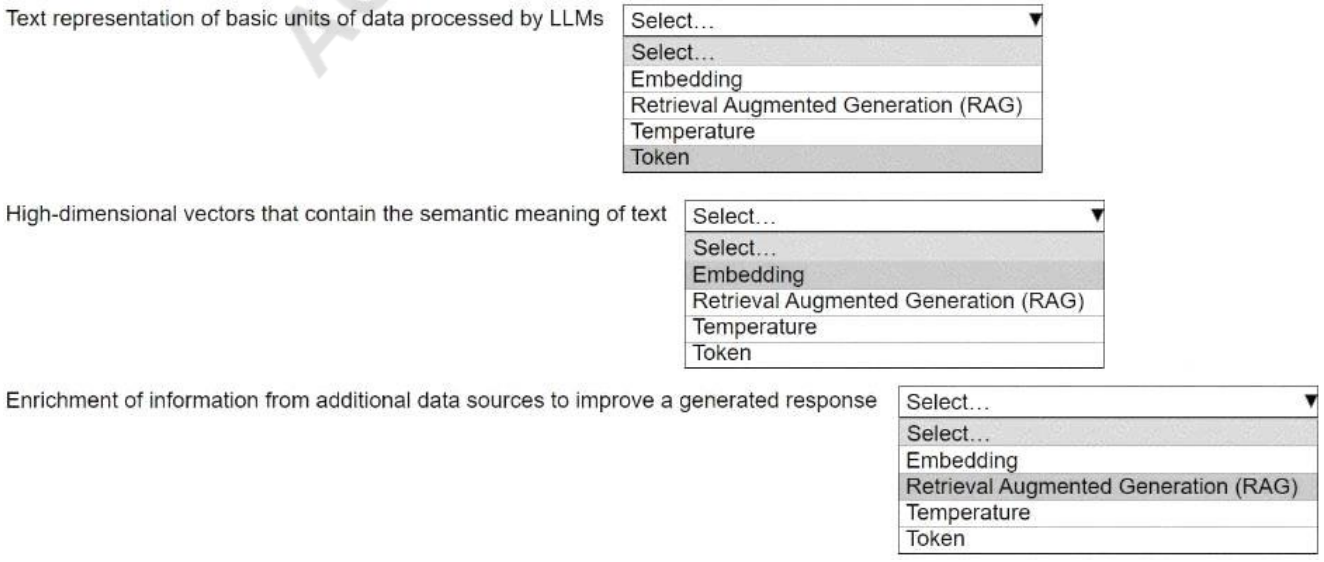
NO.50 An ML engineer is building a generative AI application on Amazon Bedrock by using large language models (LLMs).
Select the correct generative AI term from the following list for each description. Each term should be selected one time or not at all. (Select three.)
* Embedding
* Retrieval Augmented Generation (RAG)
* Temperature
* Token



Answer:



Explanation:



* Text representation of basic units of data processed by LLMs:Token
* High-dimensional vectors that contain the semantic meaning of text:Embedding
* Enrichment of information from additional data sources to improve a generated response:

Retrieval Augmented Generation (RAG)
Comprehensive Detailed Explanation
* Token:
* Description: A token represents the smallest unit of text (e.g., a word or part of a word) that an LLM processes. For example, "running" might be split into two tokens: "run" and "ing."
* Why?Tokens are the fundamental building blocks for LLM input and output processing, ensuring that the model can understand and generate text efficiently.
* Embedding:
* Description: High-dimensional vectors that encode the semantic meaning of text. These vectors are representations of words, sentences, or even paragraphs in a way that reflects their relationships and meaning.
* Why?Embeddings are essential for enabling similarity search, clustering, or any task requiring semantic understanding. They allow the model to "understand" text contextually.
* Retrieval Augmented Generation (RAG):
* Description: A technique where information is enriched or retrieved from external data sources (e.g., knowledge bases or document stores) to improve the accuracy and relevance of a model's generated responses.
* Why?RAG enhances the generative capabilities of LLMs by grounding their responses in factual and up-to-date information, reducing hallucinations in generated text.
By matching these terms to their respective descriptions, the ML engineer can effectively leverage these concepts to build robust and contextually aware generative AI applications on Amazon Bedrock.

NO.51 A company needs to host a custom ML model to perform forecast analysis. The forecast analysis will occur with predictable and sustained load during the same 2-hour period every day. Multiple invocations during the analysis period will require quick responses. The company needs AWS to manage the underlying infrastructure and any auto scaling activities.
Which solution will meet these requirements?

A. Schedule an Amazon SageMaker batch transform job by using AWS Lambda.

B. Configure an Auto Scaling group of Amazon EC2 instances to use scheduled scaling.

C. Use Amazon SageMaker Serverless Inference with provisioned concurrency.

D. Run the model on an Amazon Elastic Kubernetes Service (Amazon EKS) cluster on Amazon EC2 with pod auto scaling.

Answer: C
Explanation:
SageMaker Serverless Inference is ideal for workloads with predictable, intermittent demand. By enabling provisioned concurrency, the model can handle multiple invocations quickly during the high-demand 2-hour period. AWS manages the underlying infrastructure and scaling, ensuring the solution meets performance requirements with minimal operational overhead. This approach is cost-effective since it scales down when not in use.

NO.52 A company is building a deep learning model on Amazon SageMaker. The company uses a large amount of data as the training dataset. The company needs to optimize the model's hyperparameters to minimize the loss function on the validation dataset.
Which hyperparameter tuning strategy will accomplish this goal with the LEAST computation time?

A. Hyperbaric!

B. Grid search

C. Bayesian optimization

D. Random search

Answer: A

Explanation:

Hyperband is a hyperparameter tuning strategy designed to minimize computation time by adaptively allocating resources to promising configurations and terminating underperforming ones early. It efficiently balances exploration and exploitation, making it ideal for large datasets and deep learning models where training can be computationally expensive.

NO.53 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 Retrieval.

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 Retrieval.

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 Retrieval.

Answer: C

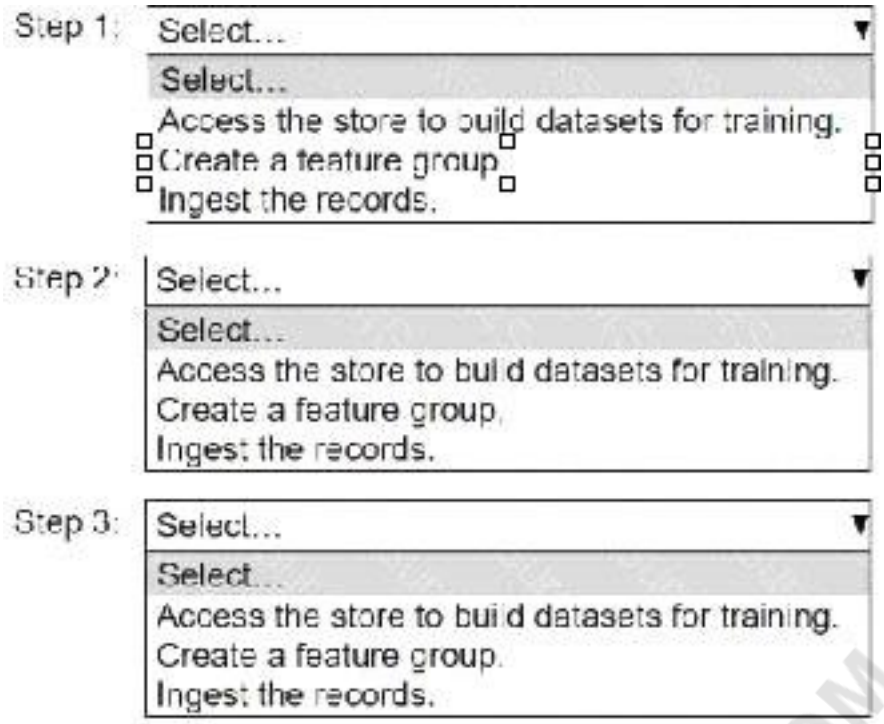
Explanation:

Partitioning the time-series data by date prefix in the S3 bucket significantly improves query performance in Amazon Athena by reducing the amount of data that needs to be scanned during queries. This allows the ML engineers to efficiently analyze trends over specific time periods, such as the past 3 days. Applying S3 Lifecycle policies to archive partitions older than 30 days to S3 Glacier FlexibleRetrieval ensures cost- effective data retention and storage management while maintaining high performance for recent data retrieval.

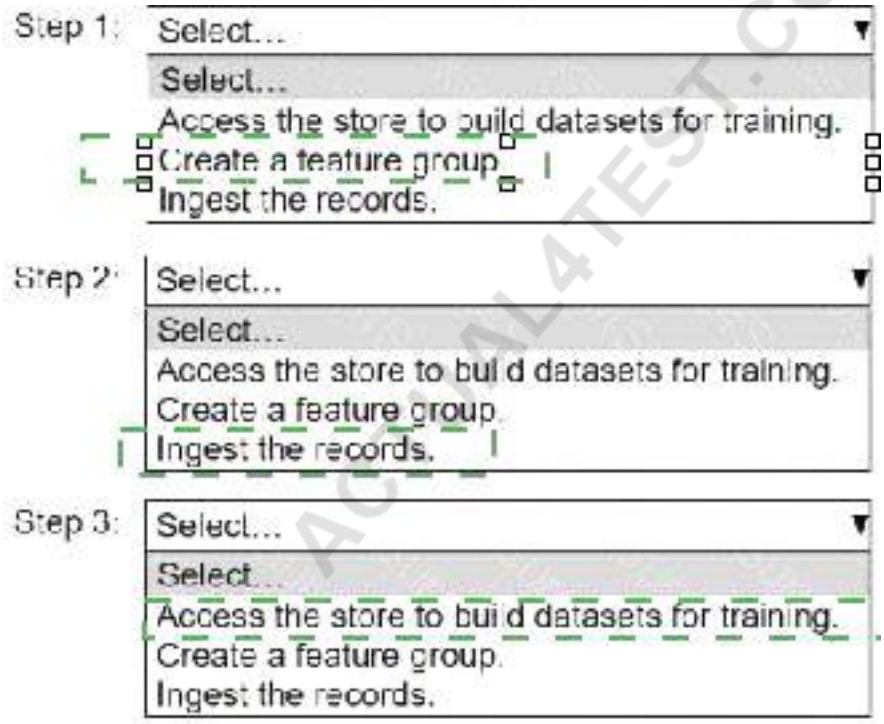
NO.54 An ML engineer needs to use Amazon SageMaker Feature Store to create and manage features to train a model.

Select and order the steps from the following list to create and use the features in Feature Store. Each step should be selected one time. (Select and order three.)

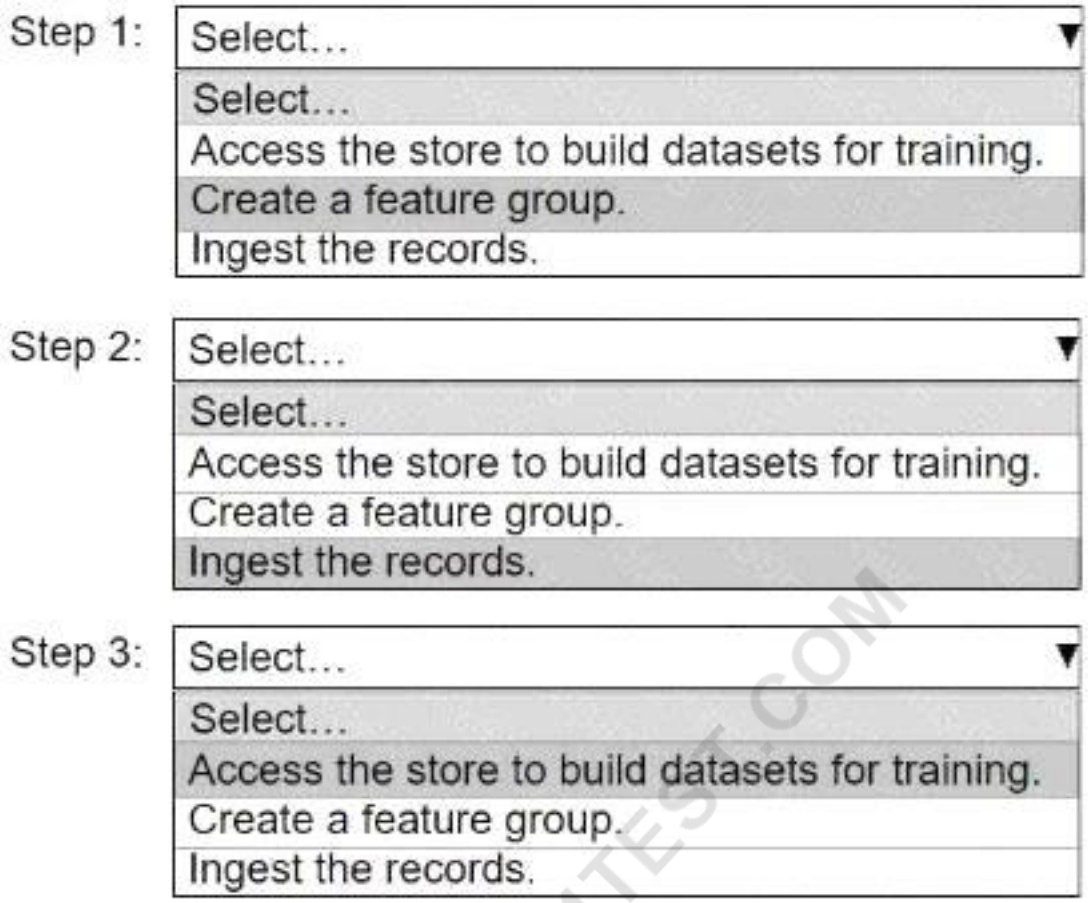
* Access the store to build datasets for training.
* Create a feature group.
* Ingest the records.



Answer:



Explanation:



Step 1: Create a feature group.Step 2: Ingest the records.Step 3: Access the store to build datasets for training.

* Step 1: Create a Feature Group
* Why?A feature group is the foundational unit in SageMaker Feature Store, where features are defined, stored, and organized. Creating a feature group specifies the schema (name, data type) for the features and the primary keys for data identification.
* How?Use the SageMaker Python SDK or AWS CLI to define the feature group by specifying its name, schema, and S3 storage location for offline access.
* Step 2: Ingest the Records
* Why?After creating the feature group, the raw data must be ingested into the Feature Store. This step populates the feature group with data, making it available for both real-time and offline use.
* How?Use the SageMaker SDK or AWS CLI to batch-ingest historical data or stream new records into the feature group. Ensure the records conform to the feature group schema.
* Step 3: Access the Store to Build Datasets for Training
* Why?Once the features are stored, they can be accessed to create training datasets. These datasets combine relevant features into a single format for machine learning model training.
* How?Use the SageMaker Python SDK to query the offline store or retrieve real-time features using the online store API. The offline store is typically used for batch training, while the online store is used for inference.

Order Summary:

* Create a feature group.
* Ingest the records.
* Access the store to build datasets for training.

This process ensures the features are properly managed, ingested, and accessible for model training

using Amazon SageMaker Feature Store.

NO.55 A company wants to predict the success of advertising campaigns by considering the color scheme of each advertisement. An ML engineer is preparing data for a neural network model. The dataset includes color information as categorical data.
Which technique for feature engineering should the ML engineer use for the model?
A. Apply label encoding to the color categories. Automatically assign each color a unique integer.
B. Implement padding to ensure that all color feature vectors have the same length.
C. Perform dimensionality reduction on the color categories.
D. One-hot encode the color categories to transform the color scheme feature into a binary matrix.
Answer: D
Explanation:
One-hot encodingis the appropriate technique for transforming categorical data, such as color information, into a format suitable for input to a neural network. This technique creates a binary vector representation where each unique category (color) is represented as a separate binary column, ensuring that the model does not infer ordinal relationships between categories. This approach preserves the categorical nature of the data and avoids introducing unintended biases.

NO.56 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.
Answer: C
Explanation:
AWS Lake Formation provides fine-grained access control and simplifies data governance for data lakes. By configuring Lake Formation tags to map ML engineers to their specific campaigns, you can restrict access to both structured and unstructured data in the data lake. This method is operationally efficient, as it centralizes access control management within Lake Formation and ensures consistency across Amazon Athena and S3 bucket access without requiring manual updates to policies or DynamoDB-based custom logic.

NO.57 A company has trained an ML model in Amazon SageMaker. The company needs to host the model to provide inferences in a production environment.
The model must be highly available and must respond with minimum latency. The size of each

request will be between 1 KB and 3 MB. The model will receive unpredictable bursts of requests during the day. The inferences must adapt proportionally to the changes in demand.
How should the company deploy the model into production to meet these requirements?

A. Create a SageMaker real-time inference endpoint. Configure auto scaling. Configure the endpoint to present the existing model.

B. Deploy the model on an Amazon Elastic Container Service (Amazon ECS) cluster. Use ECS scheduled scaling that is based on the CPU of the ECS cluster.

C. Install SageMaker Operator on an Amazon Elastic Kubernetes Service (Amazon EKS) cluster. Deploy the model in Amazon EKS. Set horizontal pod auto scaling to scale replicas based on the memory metric.

D. Use Spot Instances with a Spot Fleet behind an Application Load Balancer (ALB) for inferences. Use the ALBRequestCountPerTarget metric as the metric for auto scaling.

Answer: A

Explanation:
Amazon SageMaker real-time inference endpoints are designed to provide low-latency predictions in production environments. They offer built-in auto scaling to handle unpredictable bursts of requests, ensuring high availability and responsiveness. This approach is fully managed, reduces operational complexity, and is optimized for the range of request sizes (1 KB to 3 MB) specified in the requirements.

NO.58 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 Resource 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.

Answer: C

Explanation:
Callback steps in Amazon SageMaker Pipelines allow you to integrate external processes, such as AWS Glue jobs, into the pipeline workflow. By using a Callback step, the SageMaker pipeline can trigger the AWS Glue workflow and pause execution until the Glue jobs complete. This approach provides seamless integration with minimal operational overhead, as it directly ties the pipeline's execution flow to the completion of the AWS Glue jobs without requiring additional orchestration tools or complex setups.

NO.59 A company is running ML models on premises by using custom Python scripts and proprietary datasets. The company is using PyTorch. 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.

Answer: B

Explanation:

SageMaker script mode allows you to bring existing custom Python scripts and run them on AWS with minimal changes. SageMaker provides prebuilt containers for ML frameworks like PyTorch, simplifying the migration process. This approach enables the company to leverage their existing Python scripts and domain knowledge while benefiting from the scalability and managed environment of SageMaker. It requires the least effort compared to building custom containers or retraining models from scratch.

NO.60 A company is gathering audio, video, and text data in various languages. The company needs to use a large language model (LLM) to summarize the gathered data that is in Spanish.

Which solution will meet these requirements in the LEAST amount of time?

A. Train and deploy a model in Amazon SageMaker to convert the data into English text. Train and deploy an LLM in SageMaker to summarize the text.

B. Use Amazon Transcribe and Amazon Translate to convert the data into English text. Use Amazon Bedrock with the Jurassic model to summarize the text.

C. Use Amazon Rekognition and Amazon Translate to convert the data into English text. Use Amazon Bedrock with the Anthropic Claude model to summarize the text.

D. Use Amazon Comprehend and Amazon Translate to convert the data into English text. Use Amazon Bedrock with the Stable Diffusion model to summarize the text.

Answer: B

Explanation:

Amazon Transcribeis well-suited for converting audio data into text, including Spanish.

Amazon Translatecan efficiently translate Spanish text into English if needed.

Amazon Bedrock, with theJurassic model, is designed for tasks like text summarization and can handle large language models (LLMs) seamlessly. This combination provides a low-code, managed solution to process audio, video, and text data with minimal time and effort.

NO.61 A company regularly receives new training data from the vendor of an ML model. The vendor delivers cleaned and prepared data to the company's Amazon S3 bucket every 3-4 days.

The company has an Amazon SageMaker pipeline to retrain the model. An ML engineer needs to implement a solution to run the pipeline when new data is uploaded to the S3 bucket.

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

A. Create an S3 Lifecycle rule to transfer the data to the SageMaker training instance and to initiate training.

B. Create an AWS Lambda function that scans the S3 bucket. Program the Lambda function to initiate the pipeline when new data is uploaded.

C. Create an Amazon EventBridge rule that has an event pattern that matches the S3 upload. Configure the pipeline as the target of the rule.

D. Use Amazon Managed Workflows for Apache Airflow (Amazon MWAA) to orchestrate the pipeline when new data is uploaded.

Answer: C

Explanation:

UsingAmazon EventBridgewith an event pattern that matches S3 upload events provides an automated, low- effort solution. When new data is uploaded to the S3 bucket, the EventBridge rule triggers the SageMaker pipeline. This approach minimizes operational overhead by eliminating the need for custom scripts or external orchestration tools while seamlessly integrating with the existing S3 and SageMaker setup.

NO.62 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.

Answer: D

Explanation:

A serverless inference endpoint in Amazon SageMaker is ideal for use cases where the model is invoked infrequently, such as running one time each night. It eliminates the cost of idle resources when the model is not in use. Setting the MaxConcurrency parameter to 1 ensures cost-efficiency while supporting the required single nightly invocation. This solution minimizes costs and matches the requirement to process a small amount of data quickly.

NO.63 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.

Answer: D

Explanation:

Amazon SageMaker batch transform is ideal for obtaining inferences from large datasets in an

asynchronous manner, as it processes data in batches rather than requiring real-time inputs. SageMaker Model Monitor allows scheduled monitoring of data quality, detecting shifts in input data characteristics, and generating alerts when changes in data quality occur.
This solution provides a fully managed, efficient way to handle both asynchronous inference and data quality monitoring with minimal operational overhead.

NO.64 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 a single data frame. Use Amazon SageMaker data labeling to prepare the data.
Answer: A
Explanation:
Amazon SageMakerData Wranglerprovides a comprehensive solution for importing, consolidating, and preparing datasets for ML. It offers tools to handle missing values, duplicates, and outliers through its built- incleansingandenrichmentfunctionalities, allowing the ML engineer to efficiently prepare the data in a single environment with minimal manual effort.

NO.65 An ML engineer is training a simple neural network model. The ML engineer tracks the performance of the model over time on a validation dataset. The model's performance improves substantially at first and then degrades after a specific number of epochs.
Which solutions will mitigate this problem? (Choose two.)
A. Enable early stopping on the model.
B. Increase dropout in the layers.
C. Increase the number of layers.
D. Increase the number of neurons.
E. Investigate and reduce the sources of model bias.
Answer: A B
Explanation:
Early stopping halts training once the performance on the validation dataset stops improving. This prevents the model from overfitting, which is likely the cause of performance degradation after a certain number of epochs.
Dropout is a regularization technique that randomly deactivates neurons during training, reducing overfitting by forcing the model to generalize better. Increasing dropout can help mitigate the problem of performance degradation due to overfitting.

NO.66 An ML engineer has developed a binary classification model outside of Amazon SageMaker. The ML engineer needs to make the model accessible to a SageMaker Canvas user for additional

tuning.
The model artifacts are stored in an Amazon S3 bucket. The ML engineer and the Canvas user are part of the same SageMaker domain.
Which combination of requirements must be met so that the ML engineer can share the model with the Canvas user? (Choose two.)

A. The ML engineer and the Canvas user must be in separate SageMaker domains.

B. The Canvas user must have permissions to access the S3 bucket where the model artifacts are stored.

C. The model must be registered in the SageMaker Model Registry.

D. The ML engineer must host the model on AWS Marketplace.

E. The ML engineer must deploy the model to a SageMaker endpoint.

Answer: B C

Explanation:
The SageMaker Canvas user needs permissions to access the Amazon S3 bucket where the model artifacts are stored to retrieve the model for use in Canvas.
Registering the model in the SageMaker Model Registry allows the model to be tracked and managed within the SageMaker ecosystem. This makes it accessible for tuning and deployment through SageMaker Canvas.
This combination ensures proper access control and integration within SageMaker, enabling the Canvas user to work with the model.

NO.67 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.

Answer: D

Explanation:
AWS Compute Optimizer analyzes the resource usage of Amazon EC2 instances, ECS services, Lambda functions, and Amazon EBS volumes. It provides actionable recommendations to optimize resource utilization and reduce costs, such as resizing instances, moving workloads to Spot Instances, or changing volume types. This solution requires the least development effort because Compute Optimizer is a managed service that automatically generates insights and recommendations based on historical usage data.

NO.68 A company has a large collection of chat recordings from customer interactions after a product release. An ML engineer needs to create an ML model to analyze the chat data. The ML engineer needs to determine the success of the product by reviewing customer sentiments about the product.

Which action should the ML engineer take to complete the evaluation in the LEAST amount of time?

A. Use Amazon Rekognition to analyze sentiments of the chat conversations.

B. Train a Naive Bayes classifier to analyze sentiments of the chat conversations.

C. Use Amazon Comprehend to analyze sentiments of the chat conversations.

D. Use random forests to classify sentiments of the chat conversations.

Answer: C

Explanation:

Amazon Comprehend is a fully managed natural language processing (NLP) service that includes a built-in sentiment analysis feature. It can quickly and efficiently analyze text data to determine whether the sentiment is positive, negative, neutral, or mixed. Using Amazon Comprehend requires minimal setup and provides accurate results without the need to train and deploy custom models, making it the fastest and most efficient solution for this task.

NO.69 An ML engineer is using a training job to fine-tune a deep learning model in Amazon SageMaker Studio. The ML engineer previously used the same pre-trained model with a similar dataset. The ML engineer expects vanishing gradient, underutilized GPU, and overfitting problems. The ML engineer needs to implement a solution to detect these issues and to react in predefined ways when the issues occur. The solution also must provide comprehensive real-time metrics during the training.

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

A. Use TensorBoard to monitor the training job. Publish the findings to an Amazon Simple Notification Service (Amazon SNS) topic. Create an AWS Lambda function to consume the findings and to initiate the predefined actions.

B. Use Amazon CloudWatch default metrics to gain insights about the training job. Use the metrics to invoke an AWS Lambda function to initiate the predefined actions.

C. Expand the metrics in Amazon CloudWatch to include the gradients in each training step. Use the metrics to invoke an AWS Lambda function to initiate the predefined actions.

D. Use SageMaker Debugger built-in rules to monitor the training job. Configure the rules to initiate the predefined actions.

Answer: D

Explanation:

SageMaker Debugger provides built-in rules to automatically detect issues like vanishing gradients, underutilized GPU, and overfitting during training jobs. It generates real-time metrics and allows users to define predefined actions that are triggered when specific issues occur. This solution minimizes operational overhead by leveraging the managed monitoring capabilities of SageMaker Debugger without requiring custom setups or extensive manual intervention.

NO.70 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.

Answer: D

Explanation:

Logging the metrics to Amazon CloudWatch allows the metrics to be tracked and monitored effectively.

CloudWatch Alarms can be configured to trigger when metrics breach a predefined threshold.

The alarm can be set to notify through Amazon Simple Notification Service (SNS), which can send email messages to the configured recipients.

This is the standard and most efficient way to achieve the desired functionality.

NO.71 An ML engineer needs to use an Amazon EMR cluster to process large volumes of data in batches. Any data loss is unacceptable.

Which instance purchasing option will meet these requirements MOST cost-effectively?

A. Run the primary node, core nodes, and task nodes on On-Demand Instances.

B. Run the primary node, core nodes, and task nodes on Spot Instances.

C. Run the primary node on an On-Demand Instance. Run the core nodes and task nodes on Spot Instances.

D. Run the primary node and core nodes on On-Demand Instances. Run the task nodes on Spot Instances.

Answer: D

Explanation:

For Amazon EMR, the primary node and core nodes handle the critical functions of the cluster, including data storage (HDFS) and processing. Running them on On-Demand Instances ensures high availability and prevents data loss, as Spot Instances can be interrupted. The task nodes, which handle additionalprocessing but do not store data, can use Spot Instances to reduce costs without compromising the cluster's resilience or data integrity. This configuration balances cost-effectiveness and reliability.

NO.72 A company has trained and deployed an ML model by using Amazon SageMaker. The company needs to implement a solution to record and monitor all the API call events for the SageMaker endpoint. The solution also must provide a notification when the number of API call events breaches a threshold.

Use SageMaker Debugger to track the inferences and to report metrics. Create a custom rule to provide a notification when the threshold is breached.

Which solution will meet these requirements?

A. Use SageMaker Debugger to track the inferences and to report metrics. Create a custom rule to provide a notification when the threshold is breached.

B. Use SageMaker Debugger to track the inferences and to report metrics. Use the tensor_variance built-in rule to provide a notification when the threshold is breached.

C. Log all the endpoint invocation API events by using AWS CloudTrail. Use an Amazon CloudWatch dashboard for monitoring. Set up a CloudWatch alarm to provide notification when the threshold is breached.

D. Add the Invocations metric to an Amazon CloudWatch dashboard for monitoring. Set up a CloudWatch alarm to provide notification when the threshold is breached.

Answer: D

Explanation:

Amazon SageMaker automatically tracks theInvocationsmetric, which represents the number of API calls made to the endpoint, inAmazon CloudWatch. By adding this metric to a CloudWatch dashboard, you can monitor the endpoint's activity in real-time. Setting up aCloudWatch alarmallows the system to send notifications whenever the API call events exceed the defined threshold, meeting both the monitoring and notification requirements efficiently.

NO.73 A company needs to create a central catalog for all the company's ML models. The models are in AWS accounts where the company developed the models initially. The models are hosted in Amazon Elastic Container Registry (Amazon ECR) repositories.

Which solution will meet these requirements?

A. Configure ECR cross-account replication for each existing ECR repository. Ensure that each model is visible in each AWS account.

B. Create a new AWS account with a new ECR repository as the central catalog. Configure ECR cross-account replication between the initial ECR repositories and the central catalog.

C. Use the Amazon SageMaker Model Registry to create a model group for models hosted in Amazon ECR. Create a new AWS account. In the new account, use the SageMaker Model Registry as the central catalog. Attach a cross-account resource policy to each model group in the initial AWS accounts.

D. Use an AWS Glue Data Catalog to store the models. Run an AWS Glue crawler to migrate the models from the ECR repositories to the Data Catalog. Configure cross-account access to the Data Catalog.

Answer: C

Explanation:

The Amazon SageMaker Model Registry is designed to manage and catalog ML models, including those hosted in Amazon ECR. By creating a model group for each model in the SageMaker Model Registry and setting up cross-account resource policies, the company can establish a central catalog in a new AWS account.

This allows all models from the initial accounts to be accessible in a unified, centralized manner for better organization, management, and governance. This solution leverages existing AWS services and ensures scalability and minimal operational overhead.

NO.74 A company is planning to create several ML prediction models. The training data is stored in Amazon S3. The entire dataset is more than 5 ## in size and consists of CSV, JSON, Apache Parquet, and simple text files.

The data must be processed in several consecutive steps. The steps include complex manipulations that can take hours to finish running. Some of the processing involves natural language processing (NLP) transformations. The entire process must be automated.

Which solution will meet these requirements?

A. Process data at each step by using Amazon SageMaker Data Wrangler. Automate the process by using Data Wrangler jobs.

B. Use Amazon SageMaker notebooks for each data processing step. Automate the process by using

Amazon EventBridge.

C. Process data at each step by using AWS Lambda functions. Automate the process by using AWS Step Functions and Amazon EventBridge.

D. Use Amazon SageMaker Pipelines to create a pipeline of data processing steps. Automate the pipeline by using Amazon EventBridge.

Answer: D

Explanation:

Amazon SageMaker Pipelines is designed for creating, automating, and managing end-to-end ML workflows, including complex data preprocessing tasks. It supports handling large datasets and can integrate with custom steps, such as NLP transformations. By combining SageMaker Pipelines with Amazon EventBridge, the entire workflow can be triggered and automated efficiently, meeting the requirements for scalability, automation, and processing complexity.

NO.75 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.

Answer: D

NO.76 A company wants to improve the sustainability of its ML operations.

Which actions will reduce the energy usage and computational resources that are associated with the company's training jobs? (Choose two.)

A. Use Amazon SageMaker Debugger to stop training jobs when non-converging conditions are detected.

B. Use Amazon SageMaker Ground Truth for data labeling.

C. Deploy models by using AWS Lambda functions.

D. Use AWS Trainium instances for training.

E. Use PyTorch or TensorFlow with the distributed training option.

Answer: A D

Explanation:

SageMaker Debuggercan identify when a training job is not converging or is stuck in a non-productive state.

By stopping these jobs early, unnecessary energy and computational resources are conserved,

improving sustainability.
AWS Trainiuminstances are purpose-built for ML training and are optimized for energy efficiency and cost- effectiveness. They use less energy per training task compared to general-purpose instances, making them a sustainable choice.

NO.77 A company uses a hybrid cloud environment. A model that is deployed on premises uses data in Amazon 53 to provide customers with a live conversational engine.
The model is using sensitive data. An ML engineer needs to implement a solution to identify and remove the sensitive data.
Which solution will meet these requirements with the LEAST operational overhead?
A. Deploy the model on Amazon SageMaker. Create a set of AWS Lambda functions to identify and remove the sensitive data.
B. Deploy the model on an Amazon Elastic Container Service (Amazon ECS) cluster that uses AWS Fargate. Create an AWS Batch job to identify and remove the sensitive data.
C. Use Amazon Macie to identify the sensitive data. Create a set of AWS Lambda functions to remove the sensitive data.
D. Use Amazon Comprehend to identify the sensitive data. Launch Amazon EC2 instances to remove the sensitive data.
Answer: C
Explanation:
Amazon Macie is a fully managed data security and privacy service that uses machine learning to discover and classify sensitive data in Amazon S3. It is purpose-built to identify sensitive data with minimal operational overhead. After identifying the sensitive data, you can use AWS Lambda functions to automate the process of removing or redacting the sensitive data, ensuring efficiency and integration with the hybrid cloud environment. This solution requires the least development effort and aligns with the requirement to handle sensitive data effectively.

NO.78 A company uses Amazon SageMaker Studio to develop an ML model. The company has a single SageMaker Studio domain. An ML engineer needs to implement a solution that provides an automated alert when SageMaker compute costs reach a specific threshold.
Which solution will meet these requirements?
A. Add resource tagging by editing the SageMaker user profile in the SageMaker domain. Configure AWS Cost Explorer to send an alert when the threshold is reached.
B. Add resource tagging by editing the SageMaker user profile in the SageMaker domain. Configure AWS Budgets to send an alert when the threshold is reached.
C. Add resource tagging by editing each user's IAM profile. Configure AWS Cost Explorer to send an alert when the threshold is reached.
D. Add resource tagging by editing each user's IAM profile. Configure AWS Budgets to send an alert when the threshold is reached.
Answer: B
Explanation:
Adding resource tagging to the SageMaker user profile enables tracking and monitoring of costs associated with specific SageMaker resources.
AWS Budgets allows setting thresholds and automated alerts for costs and usage, making it the ideal service to notify the ML engineer when compute costs reach a specified limit.

This solution is efficient and integrates seamlessly with SageMaker and AWS cost management tools.

NO.79 A financial company receives a high volume of real-time market data streams from an external provider. The streams consist of thousands of JSON records every second.
The company needs to implement a scalable solution on AWS to identify anomalous data points.
Which solution will meet these requirements with the LEAST operational overhead?

A. Ingest real-time data into Amazon Kinesis data streams. Use the built-in RANDOM_CUT_FOREST function in Amazon Managed Service for Apache Flink to process the data streams and to detect data anomalies.

B. Ingest real-time data into Amazon Kinesis data streams. Deploy an Amazon SageMaker endpoint for real-time outlier detection. Create an AWS Lambda function to detect anomalies. Use the data streams to invoke the Lambda function.

C. Ingest real-time data into Apache Kafka on Amazon EC2 instances. Deploy an Amazon SageMaker endpoint for real-time outlier detection. Create an AWS Lambda function to detect anomalies. Use the data streams to invoke the Lambda function.

D. Send real-time data to an Amazon Simple Queue Service (Amazon SQS) FIFO queue. Create an AWS Lambda function to consume the queue messages. Program the Lambda function to start an AWS Glue extract, transform, and load (ETL) job for batch processing and anomaly detection.

Answer: A

Explanation:
This solution is the most efficient and involves the least operational overhead:
Amazon Kinesis data streams efficiently handle real-time ingestion of high-volume streaming data.
Amazon Managed Service for Apache Flink provides a fully managed environment for stream processing with built-in support for RANDOM_CUT_FOREST, an algorithm designed for anomaly detection in real- time streaming data.
This approach eliminates the need for deploying and managing additional infrastructure like SageMaker endpoints, Lambda functions, or external tools, making it the most scalable and operationally simple solution.

NO.80 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.

Answer: A

Explanation:

SageMaker real-time inference is designed for low-latency, real-time use cases, such as detecting fraudulent transactions in banking applications. It eliminates the delays associated with SageMaker Asynchronous Inference, improving inference performance.
SageMaker Model Monitor provides tools to monitor deployed models for deviations in data quality, model performance, and other metrics. It can be configured to send notifications when a deviation in model quality is detected, ensuring the system remains reliable.

NO.81 Case study
An ML engineer is developing a fraud detection model on AWS. The training dataset includes transaction logs, customer profiles, and tables from an on-premises MySQL database. The transaction logs and customer profiles are stored in Amazon S3.
The dataset has a class imbalance that affects the learning of the model's algorithm. Additionally, many of the features have interdependencies. The algorithm is not capturing all the desired underlying patterns in the data.
The training dataset includes categorical data and numerical data. The ML engineer must prepare the training dataset to maximize the accuracy of the model.
Which action will meet this requirement with the LEAST operational overhead?

A. Use AWS Glue to transform the categorical data into numerical data.

B. Use AWS Glue to transform the numerical data into categorical data.

C. Use Amazon SageMaker Data Wrangler to transform the categorical data into numerical data.

D. Use Amazon SageMaker Data Wrangler to transform the numerical data into categorical data.

Answer: C

Explanation:
Preparing a training dataset that includes both categorical and numerical data is essential for maximizing the accuracy of a machine learning model. Transforming categorical data into numerical format is a critical step, as most ML algorithms require numerical input.
Why Transform Categorical Data into Numerical Data?
* Model Compatibility: Many ML algorithms cannot process categorical data directly and require numerical representations.
* Improved Performance: Proper encoding of categorical variables can enhance model accuracy and convergence speed.
Why Use Amazon SageMaker Data Wrangler?
Amazon SageMaker Data Wrangler offers a visual interface with over 300 built-in data transformations, including tools for encoding categorical variables.
Implementation Steps:
* Import Data:
* Load the dataset into SageMaker Data Wrangler from sources like Amazon S3 or on-premises databases.
* Identify Categorical Features:
* Use Data Wrangler's data type inference to detect categorical columns.
* Apply Categorical Encoding:
* Choose appropriate encoding techniques (e.g., one-hot encoding or ordinal encoding) from Data Wrangler's transformation options.
* Apply the selected transformation to convert categorical features into numerical format.
* Validate Transformations:
* Review the transformed dataset to ensure accuracy and completeness.

Advantages of Using SageMaker Data Wrangler:
* Ease of Use: Provides a user-friendly interface for data transformation without extensive coding.
* Operational Efficiency: Integrates data preparation steps, reducing the need for multiple tools and minimizing operational overhead.
* Flexibility: Supports various data sources and transformation techniques, accommodating diverse datasets.
By utilizing SageMaker Data Wrangler to transform categorical data into numerical format, the ML engineer can efficiently prepare the dataset, thereby enhancing the model's accuracy with minimal operational overhead.
References:
* Transform Data - Amazon SageMaker
* Prepare ML Data with Amazon SageMaker Data Wrangler

NO.82 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
Answer: D
Explanation:
Apache Parquet is a columnar storage file format optimized for complex and large datasets. It provides efficient reading and processing by accessing only the required columns, which reduces I/O and speeds up data handling. This makes it ideal for use with Amazon SageMaker Canvas, where minimizing processing time is important for training ML models. Parquet is also compatible with S3 and widely supported in data analytics and ML workflows.

NO.83 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 use the central model registry to manage different versions of models in the application.
Which action will meet this requirement with the LEAST operational overhead?
A. Create a separate Amazon Elastic Container Registry (Amazon ECR) repository for each model.
B. Use Amazon Elastic Container Registry (Amazon ECR) and unique tags for each model version.
C. Use the SageMaker Model Registry and model groups to catalogthe models.
D. Use the SageMaker Model Registry and unique tags for each model version.
Answer: C
Explanation:
Amazon SageMaker Model Registry is a feature designed to manage machine learning (ML) models

throughout their lifecycle. It allows users to catalog, version, and deploy models systematically, ensuring efficient model governance and management.
Key Features of SageMaker Model Registry:
* Centralized Cataloging: Organizes models intoModel Groups, each containing multiple versions.
* Version Control: Maintains a history of model iterations, making it easier to track changes.
* Metadata Association: Attach metadata such as training metrics and performance evaluations to models.
* Approval Status Management: Allows setting statuses like PendingManualApproval or Approved to ensure only vetted models are deployed.
* Seamless Deployment: Direct integration with SageMaker deployment capabilities for real-time inference or batch processing.
Implementation Steps:
* Create a Model Group: Organize related models into groups to simplify management and versioning.
* Register Model Versions: Each model iteration is registered as a version within a specific Model Group.
* Set Approval Status: Assign approval statuses to models before deploying them to ensure quality control.
* Deploy the Model: Use SageMaker endpoints for deployment once the model is approved.
Benefits:
* Centralized Management: Provides a unified platform to manage models efficiently.
* Streamlined Deployment: Facilitates smooth transitions from development to production.
* Governance and Compliance: Supports metadata association and approval processes.
By leveraging the SageMaker Model Registry, the company can ensure organized management of models, version control, and efficient deployment workflows with minimal operational overhead.
References:
* AWS Documentation: SageMaker Model Registry
* AWS Blog: Model Registry Features and Usage

NO.84 A company is using Amazon SageMaker and millions of files to train an ML model. Each file is several megabytes in size. The files are stored in an Amazon S3 bucket. The company needs to improve training performance.
Which solution will meet these requirements in the LEAST amount of time?

A. Transfer the data to a new S3 bucket that provides S3 Express One Zone storage. Adjust the training job to use the new S3 bucket.

B. Create an Amazon FSx for Lustre file system. Link the file system to the existing S3 bucket. Adjust the training job to read from the file system.

C. Create an Amazon Elastic File System (Amazon EFS) file system. Transfer the existing data to the file system. Adjust the training job to read from the file system.

D. Create an Amazon ElastiCache (Redis OSS) cluster. Link the Redis OSS cluster to the existing S3 bucket. Stream the data from the Redis OSS cluster directly to the training job.

Answer: B
Explanation:
Amazon FSx for Lustre is designed for high-performance workloads like ML training. It provides fast, low- latency access to data by linking directly to the existing S3 bucket and caching frequently accessed files locally. This significantly improves training performance compared to directly accessing

millions of files from S3. It requires minimal changes to the training job and avoids the overhead of transferring or restructuring data, making it the fastest and most efficient solution.

NO.85 An ML engineer needs to use AWS services to identify and extract meaningful unique keywords from documents.
Which solution will meet these requirements with the LEAST operational overhead?
A. Use the Natural Language Toolkit (NLTK) library on Amazon EC2 instances for text pre-processing. Use the Latent Dirichlet Allocation (LDA) algorithm to identify and extract relevant keywords.
B. Use Amazon SageMaker and the BlazingText algorithm. Apply custom pre-processing steps for stemming and removal of stop words. Calculate term frequency-inverse document frequency (TF-IDF) scores to identify and extract relevant keywords.
C. Store the documents in an Amazon S3 bucket. Create AWS Lambda functions to process the documents and to run Python scripts for stemming and removal of stop words. Use bigram and trigram techniques to identify and extract relevant keywords.
D. Use Amazon Comprehend custom entity recognition and key phrase extraction to identify and extract relevant keywords.
Answer: D
Explanation:
Amazon Comprehend provides pre-built functionality for key phrase extraction and can identify meaningful keywords from documents with minimal setup or operational overhead. It eliminates the need for manual preprocessing, stemming, or stop-word removal and does not require custom model development or infrastructure management. This makes it the most efficient and low-maintenance solution for the task.