Amazon.MLA-C01.v2026-04-16.q120

Exam Code:	MLA-C01
Exam Name:	AWS Certified Machine Learning Engineer - Associate
Certification Provider:	Amazon
Free Question Number:	120
Version:	v2026-04-16
# of views:	293
# of Questions views:	1200
https://www.freeqas.com/qa/Amazon/MLA-C01/Amazon.MLA-C01.v2026-04-16.q120.html	

NEW QUESTION: 1
A company is developing a generative AI conversational interface to assist customers with payments. The company wants to use an ML solution to detect customer intent. The company does not have training data to train a model.
Which solution will meet these requirements?
A. Fine-tune a sequence-to-sequence (seq2seq) algorithm in Amazon SageMaker JumpStart.
B. Use an LLM from Amazon Bedrock with zero-shot learning.
C. Use the Amazon Comprehend DetectEntities API.
D. Run an LLM from Amazon Bedrock on Amazon EC2 instances.
Answer: (SHOW ANSWER)
The key requirement in this scenario is detecting customer intent without having any training data. According to AWS Machine Learning and Generative AI documentation, zero-shot learning is specifically designed for situations where labeled training data is unavailable. Zero-shot learning allows a pre-trained large language model (LLM) to perform tasks it has not been explicitly trained on by leveraging its general knowledge and language understanding.
Amazon Bedrock provides fully managed access to foundation models (FMs) and LLMs that support zero- shot and few-shot learning. By using an LLM from Amazon Bedrock, the company can directly infer customer intent from natural language inputs without building, training, or fine-tuning a custom model. This approach is ideal for conversational interfaces where rapid deployment and scalability are required.
Option A is incorrect because fine-tuning a sequence-to-sequence (seq2seq) model in Amazon SageMaker JumpStart still requires labeled training data. Since the company explicitly does not have training data, this option does not meet the requirement.
Option C is also incorrect because the Amazon Comprehend DetectEntities API is designed for named entity recognition (NER), such as detecting names, dates, locations, or monetary values. It does not perform intent detection and is not suitable for conversational AI intent classification.
Option D is partially misleading. While it is technically possible to run an LLM on Amazon EC2, this does not inherently solve the problem of intent detection without training data. Additionally, Amazon Bedrock already abstracts infrastructure management, scaling, and model hosting, making direct EC2 deployment unnecessary and less efficient.
Therefore, using an LLM from Amazon Bedrock with zero-shot learning is the most appropriate, scalable, and AWS-recommended solution for intent detection without training data.

NEW QUESTION: 2
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 (LEAVE A REPLY)
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.

NEW QUESTION: 3
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. 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.
B. Manually import and merge the datasets. Consolidate the datasets into a single data frame. Use Amazon SageMaker data labeling to prepare the data.
C. 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.
D. 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.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 4
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. Recreate the training dataset by using the Data Wrangler resize image transform. Crop all images to the same size.
B. Recreate the training dataset by using the Data Wrangler enhance image contrast transform.

Specify the Gamma contrast option.
C. Recreate the training dataset by using the Data Wrangler corrupt image transform. Specify the impulse noise option.
D. Collect more images from all the cameras. Use Data Wrangler to prepare a new training dataset.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 5
A company is developing a customer support AI assistant by using an Amazon Bedrock Retrieval Augmented Generation (RAG) pipeline. The AI assistant retrieves articles from a knowledge base stored in Amazon S3.
The company uses Amazon OpenSearch Service to index the knowledge base. The AI assistant uses an Amazon Bedrock Titan Embeddings model for vector search.
The company wants to improve the relevance of the retrieved articles to improve the quality of the AI assistant's answers.
Which solution will meet these requirements?
A. Use auto-summarization on the retrieved articles by using Amazon SageMaker JumpStart.
B. Use a reranker model before passing the articles to the foundation model (FM).
C. Use Amazon Athena to pre-filter the articles based on metadata before retrieval.
D. Use Amazon Bedrock Provisioned Throughput to process queries more efficiently.
Answer: B (LEAVE A REPLY)
In a Retrieval Augmented Generation (RAG) architecture, retrieval quality directly impacts response accuracy. AWS documentation for Bedrock and OpenSearch highlights the use of reranker models to improve relevance after initial vector search retrieval.
Vector search retrieves documents based on embedding similarity, but the top results are not always the most contextually relevant. A reranker model evaluates the retrieved documents against the user query and reorders them based on semantic relevance before sending them to the foundation model.
Option A improves readability but does not improve retrieval relevance. Option C filters data before retrieval, which can reduce recall. Option D improves performance, not relevance.
AWS explicitly recommends reranking as a best practice for improving answer quality in RAG systems.
Therefore, Option B is the correct solution.

NEW QUESTION: 6
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 Amazon Comprehend custom entity recognition and key phrase extraction to identify and extract relevant keywords.
B. 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.
C. 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.
D. 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.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 7
A company has an existing Amazon SageMaker model (v1) on a production endpoint. The company develops a new model version (v2) and needs to test v2 in production before substituting v2 for v1.
The company needs to implement a solution to minimize the risk of v2 generating incorrect output in production. The solution must prevent any disruption of production traffic during the change to v2.
Which solution will meet these requirements?
A. Create a second production variant for v2. Assign 1% of the traffic to v2 and 99% of the traffic to v1. Collect all the output of v2 in an Amazon S3 bucket. If v2 performs as expected, switch all the traffic to v2.

B. Create a second production variant for v2. Assign 10% of the traffic to v2 and 90% of the traffic to v1. Collect all the output of v2 in an Amazon S3 bucket. If v2 performs as expected, switch all the traffic to v2.
C. Deploy v2 to a new endpoint. Turn on data capturing for the production endpoint. Write a script to pass 100% of input data to v2. If v2 performs as expected, deactivate the v1 endpoint and direct the traffic to v2.
D. Deploy v2 into a shadow variant that samples 100% of the inference requests. Collect all the output in an Amazon S3 bucket. If v2 performs as expected, promote v2 to production.
Answer: D (LEAVE A REPLY)
A shadow variant allows the new model (v2) to receive a copy of 100% of production traffic while only v1's outputs are returned to users. This enables safe side-by-side evaluation of v2 without impacting production responses, minimizing risk and ensuring no disruption of live traffic until v2 is validated and promoted.

NEW QUESTION: 8
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 AWS Glue DataBrew built-in features to oversample the minority class.
B. Use the Amazon SageMaker Data Wrangler balance data operation to oversample the minority class.
C. Use Amazon Athena to identify patterns that contribute to the imbalance. Adjust the dataset accordingly.
D. Use Amazon SageMaker Studio Classic built-in algorithms to process the imbalanced dataset.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 9
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: (SHOW ANSWER)
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.

NEW QUESTION: 10
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 random forests to classify sentiments of the chat conversations.

B. Use Amazon Rekognition to analyze sentiments of the chat conversations.
C. Use Amazon Comprehend to analyze sentiments of the chat conversations.
D. Train a Naive Bayes classifier to analyze sentiments of the chat conversations.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 11
An ML engineer is building a model to predict house and apartment prices. The model uses three features:
Square Meters, Price, and Age of Building. The dataset has 10,000 data rows. The data includes data points for one large mansion and one extremely small apartment.
The ML engineer must perform preprocessing on the dataset to ensure that the model produces accurate predictions for the typical house or apartment.
Which solution will meet these requirements?
A. Remove the outliers and perform a log transformation on the Square Meters variable.
B. Keep the outliers and perform normalization on the Square Meters variable.
C. Remove the outliers and perform one-hot encoding on the Square Meters variable.
D. Keep the outliers and perform one-hot encoding on the Square Meters variable.
Answer: A (LEAVE A REPLY)
In regression problems such as house price prediction, extreme values can significantly distort model learning.
In this dataset, the presence of a large mansion and an extremely small apartment represents clear outliers in the Square Meters feature. According to AWS Machine Learning best practices, outliers can disproportionately influence loss functions (such as mean squared error), leading to poor predictions for the majority of typical data points.
Removing these outliers helps the model focus on learning patterns that apply to the majority of houses and apartments, which aligns with the requirement to produce accurate predictions for typical properties. After removing outliers, applying a log transformation to the Square Meters feature further improves model performance by reducing skewness and stabilizing variance. Log transformations are commonly recommended in AWS and general ML documentation when numerical features span multiple orders of magnitude.
Option B is incorrect because normalization alone does not address the undue influence of extreme outliers.
Option C and D are incorrect because one-hot encoding is intended for categorical variables, not continuous numerical features such as square meters.
Therefore, removing outliers and applying a log transformation is the most statistically sound preprocessing approach.

NEW QUESTION: 12
A company plans to use Amazon SageMaker AI to build image classification models. The company has 6 TB of training data stored on Amazon FSx for NetApp ONTAP. The file system is in the same VPC as SageMaker AI.
An ML engineer must make the training data accessible to SageMaker AI training jobs.
Which solution will meet these requirements?
A. Mount the FSx for ONTAP file system as a volume to the SageMaker AI instance.
B. Create an Amazon S3 bucket and use Mountpoint for Amazon S3 to link the bucket to FSx for ONTAP.
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: (SHOW ANSWER)
Amazon SageMaker supports direct file system access for training jobs through Amazon FSx. AWS documentation states that FSx for NetApp ONTAP file systems can be mounted directly to SageMaker training instances when they are located in the same VPC.
Mounting the FSx file system allows SageMaker to stream large datasets efficiently without copying data into Amazon S3. This is ideal for very large datasets such as 6 TB of image data and avoids unnecessary storage duplication.
Options involving SageMaker Data Wrangler are intended for data preparation and exploration, not large- scale training data access. Mountpoint for Amazon S3 is not required and introduces additional complexity.

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

NEW QUESTION: 13
An ML engineer is building a logistic regression model to predict customer churn for subscription services.
The dataset contains two string variables: location and job_seniority_level.
The location variable has 3 distinct values, and the job_seniority_level variable has over 10 distinct values.
The ML engineer must perform preprocessing on the variables.
Which solution will meet this requirement?
A. Apply tokenization to location. Apply ordinal encoding to job_seniority_level.
B. Apply one-hot encoding to location. Apply ordinal encoding to job_seniority_level.
C. Apply binning to location. Apply standard scaling to job_seniority_level.
D. Apply one-hot encoding to location. Apply standard scaling to job_seniority_level.
Answer: B (LEAVE A REPLY)
Logistic regression requires numeric input features and is sensitive to how categorical variables are encoded.
AWS feature engineering best practices recommend one-hot encoding for low-cardinality categorical variables with no inherent order and ordinal encoding for categorical variables with a meaningful order.
The location feature has only three distinct values and no ordinal relationship, making one-hot encoding the most appropriate method. This prevents the model from inferring a false numerical relationship between locations.
The job_seniority_level feature typically has an inherent order (for example: junior, mid-level, senior, lead).
Even with more than 10 categories, ordinal encoding preserves this natural hierarchy while keeping the feature dimensionality manageable.
Tokenization is used for unstructured text, not structured categorical variables. Standard scaling applies only to continuous numeric features and is not suitable for categorical string variables.
AWS documentation explicitly highlights using one-hot encoding for nominal features and ordinal encoding for ordered categorical features when preparing data for linear models such as logistic regression.
Therefore, Option B is the correct and AWS-aligned solution.

NEW QUESTION: 14
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 (LEAVE A REPLY)
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.

NEW QUESTION: 15

A company uses a training job on Amazon SageMaker AI to train a neural network. The job first trains a model and then evaluates the model's performance ag test dataset. The company uses the results from the evaluation phase to decide if the trained model will go to production.

The training phase takes too long. The company needs solutions that can shorten training time without decreasing the model's final performance.

Select the correct solutions from the following list to meet the requirements for each description. Select each solution one time or not at all. (Select THREE.)

. Change the epoch count.

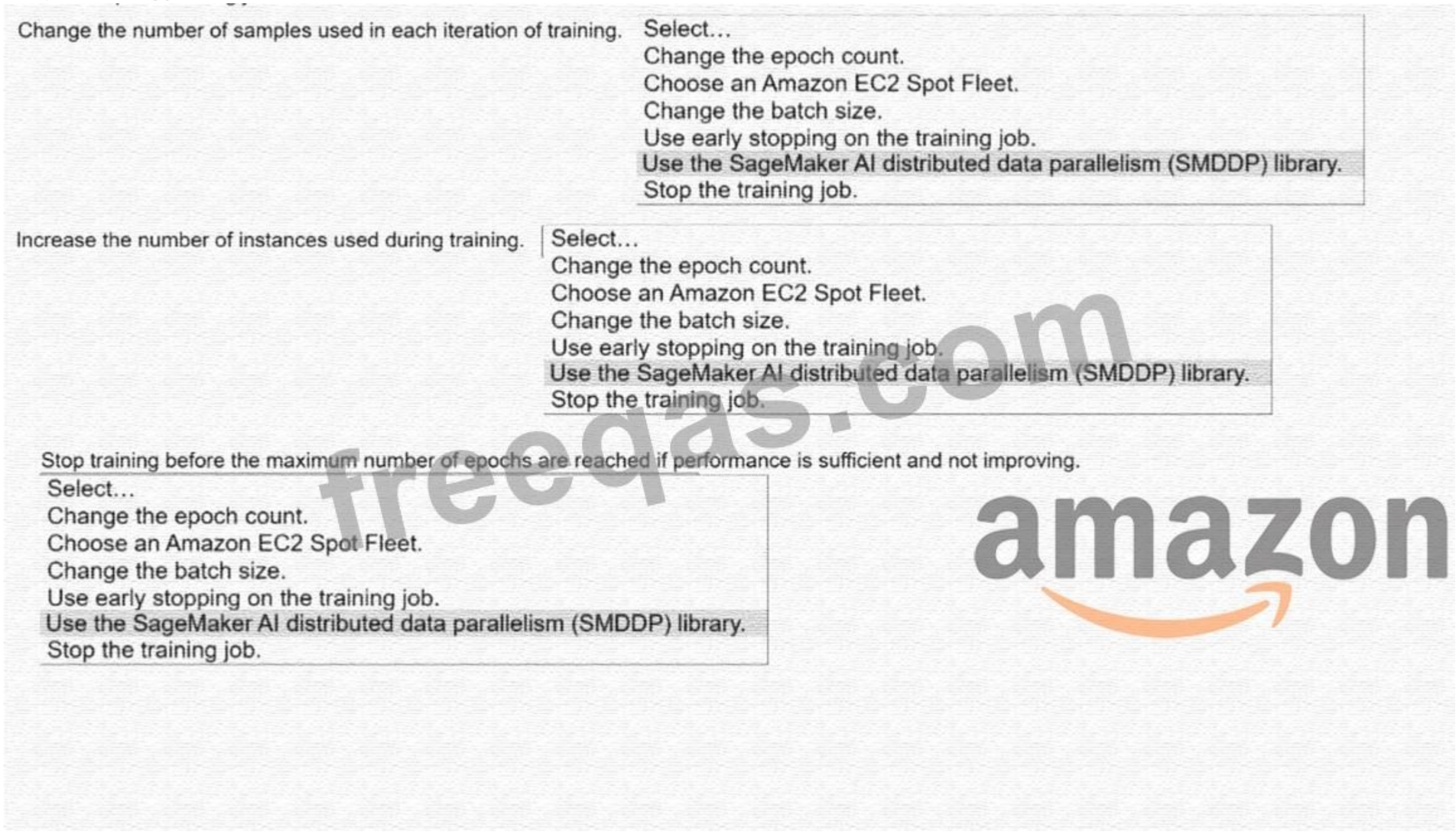
. Choose an Amazon EC2 Spot Fleet.

Change the batch size.

. Use early stopping on the training job.

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

. Stop the training job.



Answer:

Change the number of samples used in each iteration of training.

Select...
Change the epoch count.
Choose an Amazon EC2 Spot Fleet.
Change the batch size.
Use early stopping on the training job.
Use the SageMaker AI distributed data parallelism (SMDDP) library.
Stop the training job.

Increase the number of instances used during training.

Select...
Change the epoch count.
Choose an Amazon EC2 Spot Fleet.
Change the batch size.
Use early stopping on the training job.
Use the SageMaker AI distributed data parallelism (SMDDP) library.
Stop the training job.

Stop training before the maximum number of epochs are reached if performance is sufficient and not improving.

Select...
Change the epoch count.
Choose an Amazon EC2 Spot Fleet.
Change the batch size.
Use early stopping on the training job.
Use the SageMaker AI distributed data parallelism (SMDDP) library.
Stop the training job.

Explanation:
Change the number of samples used in each iteration of training
Correct selection:
Change the batch size
Why:
Increasing the batch size reduces the number of iterations per epoch, which can significantly shorten training time while maintaining model quality when tuned appropriately. AWS explicitly recommends batch size tuning as a primary performance optimization.

Increase the number of instances used during training
Correct selection:
Use the SageMaker AI distributed data parallelism (SMDDP) library
Why:
SMDDP is designed to efficiently distribute training data across multiple GPU instances with optimized gradient synchronization. This accelerates training without affecting model convergence or accuracy, unlike naive scaling approaches.
Stop training before the maximum number of epochs are reached if performance is sufficient and not improving Correct selection:
Use early stopping on the training job
Why:
Early stopping automatically terminates training when validation metrics stop improving. AWS recommends this to reduce wasted compute time while preserving optimal model performance.

NEW QUESTION: 16
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. Decrease the learning rate.
B. Increase the size of the test set.
C. Increase the learning rate.
D. Introduce early stopping.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 17
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: (SHOW ANSWER)
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.

NEW QUESTION: 18
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 (LEAVE A REPLY)
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.

NEW QUESTION: 19
An ML engineer is configuring auto scaling for an inference component of a model that runs behind an Amazon SageMaker AI endpoint. The ML engineer configures SageMaker AI auto scaling with a target tracking scaling policy set to 100 invocations per model per minute. The SageMaker AI endpoint scales appropriately during normal business hours. However, the ML engineer notices that at the start of each business day, there are zero instances available to handle requests, which causes delays in processing.
The ML engineer must ensure that the SageMaker AI endpoint can handle incoming requests at the start of each business day.
Which solution will meet this requirement?
A. Reduce the SageMaker AI auto scaling cooldown period to the minimum supported value. Add an auto scaling lifecycle hook to scale the SageMaker AI instances.
B. Change the target metric to CPU utilization.
C. Modify the scaling policy target value to one.
D. Apply a step scaling policy that scales based on an Amazon CloudWatch alarm. Apply a second CloudWatch alarm and scaling policy to scale the minimum number of instances from zero to one at the start of each business day.
Answer: (SHOW ANSWER)
This issue occurs because target tracking auto scaling allows the endpoint to scale down to zero, and scaling up only happens after traffic arrives. At the start of the business day, no instances are running, so the first requests experience cold-start delays.
AWS documentation for Amazon SageMaker recommends using scheduled or step scaling policies when predictable traffic patterns exist. In this case, business hours are predictable, so the best practice is to proactively scale the endpoint before traffic arrives.
Option D correctly uses Amazon CloudWatch alarms with step scaling to increase the minimum instance count from zero to one at the start of the business day. This ensures at least one warm instance is ready to handle requests immediately, eliminating startup latency.
Options A, B, and C do not guarantee instance availability before traffic begins. Cooldown tuning and metric changes only react after load is detected.
Therefore, scheduled step scaling using CloudWatch alarms is the correct solution.

NEW QUESTION: 20
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. JSON files compressed with gzip

B. CSV files compressed with Snappy
C. JSON objects in JSONL format
D. Apache Parquet files
Answer: (SHOW ANSWER)

NEW QUESTION: 21
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. Increase the buffer interval of the Firehose stream from 60 seconds to 120 seconds.
B. Use zero buffering in the Firehose stream. Tune the batch size that is used in the PutRecordBatch operation.
C. Replace the Firehose stream with an Amazon Simple Queue Service (Amazon SQS) queue.
D. Replace the Firehose stream with an AWS DataSync task. Configure the task with enhanced fan- out consumers.
Answer: (SHOW ANSWER)

NEW QUESTION: 22
A term frequency-inverse document frequency (tf-idf) matrix using both unigrams and bigrams is built from a text corpus consisting of the following two sentences:
1. Please call the number below.
2. Please do not call us.
What are the dimensions of the tf-idf matrix?
A. (2, 16)
B. (2, 8)
C. (2, 10)
D. (8, 10)
Answer: (SHOW ANSWER)
There are 2 sentences, 8 unique unigrams, and 8 unique bigrams, so the result would be (2,16).
The phrases are "Please call the number below" and "Please do not call us." Each word individually (unigram) is "Please," "call," "the," "number," "below," "do," "not," and "us." The unique bigrams are "Please call," "call the," "the number," "number below," "Please do," "do not,"
"not call," and "call us."

NEW QUESTION: 23
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: (SHOW ANSWER)
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.

NEW QUESTION: 24
A company is planning to use an Amazon SageMaker prebuilt algorithm to create a recommendation model. The algorithm must be able to make predictions on high-dimensional sparse data. Which SageMaker algorithm should the company choose for the recommendation model?
A. K-nearest neighbors (k-NN)
B. Factorization Machines
C. Principal component analysis (PCA)
D. Sequence-to-Sequence (seq2seq)
Answer: (SHOW ANSWER)
The Factorization Machines algorithm in SageMaker is specifically designed for recommendation systems and works well with high-dimensional sparse data such as user-item interactions. It efficiently models variable interactions and is the best choice for building a recommendation model in this scenario.

NEW QUESTION: 25
A company uses an ML model to recommend videos to users. The model is deployed on Amazon SageMaker AI. The model performed well initially after deployment, but the model's performance has degraded over time.
Which solution can the company use to identify model drift in the future?
A. Create a monitoring job in SageMaker Model Monitor. Then create a baseline from the training dataset.
B. Create a baseline from the training dataset. Then create a monitoring job in SageMaker Model Monitor.
C. Create a baseline by using a built-in rule in SageMaker Clarify. Monitor the drift in Amazon CloudWatch.
D. Retrain the model on new data. Compare the retrained model's performance to the original model's performance.
Answer: B (LEAVE A REPLY)
AWS recommends Amazon SageMaker Model Monitor for detecting data drift and model drift in deployed models. Model Monitor works by comparing live inference data against a baseline, which must first be created from the training dataset.
AWS documentation clearly specifies the required order:
* Create a baseline using training data statistics
* Create a monitoring schedule to compare incoming data against the baseline Option A reverses this order and is therefore incorrect. Option C is incorrect because SageMaker Clarify focuses on bias and explainability, not ongoing drift detection. Option D is reactive and does not provide continuous monitoring.
Model Monitor integrates with Amazon CloudWatch, enabling automated alerts and downstream retraining workflows. This proactive approach allows companies to detect degradation early and maintain model quality.
Therefore, Option B is the correct and AWS-verified answer.

NEW QUESTION: 26
A travel company wants to create an ML model to recommend the next airport destination for its users. The company has collected millions of data records about user location, recent search history on the company's website, and 2,000 available airports. The data has several categorical features with a target column that is expected to have a high-dimensional sparse matrix.
The company needs to use Amazon SageMaker AI built-in algorithms for the model. An ML engineer converts the categorical features by using one-hot encoding.
Which algorithm should the ML engineer implement to meet these requirements?
A. Use the CatBoost algorithm to recommend the next airport destination.

B. Use the DeepAR forecasting algorithm to recommend the next airport destination.
C. Use the Factorization Machines algorithm to recommend the next airport destination.
D. Use the k-means algorithm to cluster users into groups and map each group to the next airport destination.
Answer: C (LEAVE A REPLY)
This problem describes a recommendation system with millions of records, many categorical variables, and a high-dimensional sparse feature space created by one-hot encoding. AWS documentation explicitly recommends Amazon SageMaker Factorization Machines (FM) for such use cases.
Factorization Machines are designed to handle sparse datasets efficiently and to model interactions between categorical features without explicitly enumerating all feature combinations. This capability makes FM particularly well-suited for recommendation problems such as predicting user-item interactions, including destination recommendations.
With 2,000 possible airport destinations, the target space is large and sparse. One-hot encoding further increases sparsity. Factorization Machines address this challenge by learning latent factors that capture relationships between features, even when many feature combinations are rarely observed.
Option A (CatBoost) is not an Amazon SageMaker built-in algorithm and therefore does not meet the requirement. Option B (DeepAR) is a time-series forecasting algorithm, not intended for recommendation or classification problems. Option D (k-means) is an unsupervised clustering algorithm and cannot directly predict a specific destination label.
AWS documentation explicitly lists recommendation systems and click prediction as primary use cases for the SageMaker Factorization Machines algorithm.
Therefore, Option C is the correct and AWS-verified choice.

NEW QUESTION: 27
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 (LEAVE A REPLY)
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.

NEW QUESTION: 28
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 (LEAVE A REPLY)
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.

NEW QUESTION: 29
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 (LEAVE A REPLY)
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.

NEW QUESTION: 30
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 SageMaker Data Wrangler 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 AWS Batch to automatically detect the anomalies. Use Amazon QuickSight to visualize the result.
D. Use Amazon Athena to automatically detect the anomalies and to visualize the result.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 31
An ML engineer is developing a classification model. The ML engineer needs to use custom libraries in processing jobs, training jobs, and pipelines in Amazon SageMaker AI.
Which solution will provide this functionality with the LEAST implementation effort?
A. Manually install the libraries in the SageMaker AI containers.
B. Build a custom Docker container that includes the required libraries. Host the container in Amazon Elastic Container Registry (Amazon ECR). Use the ECR image in the SageMaker AI jobs and pipelines.
C. Use a SageMaker AI notebook instance and install libraries at startup.
D. Run code externally on Amazon EC2 and import results into SageMaker AI.
Answer: B (LEAVE A REPLY)
AWS documentation strongly recommends using custom Docker containers when ML workloads require consistent access to custom dependencies across processing jobs, training jobs, and pipelines. By building a single Docker image that contains all required libraries and hosting it in Amazon ECR, the ML engineer ensures that every SageMaker job uses the same runtime environment. This approach eliminates the need for repetitive installation steps and avoids environment drift.

Manually installing libraries in managed containers is error-prone and not reusable across jobs. Notebook instances are not designed to host production jobs and pipelines. Running code externally breaks the SageMaker workflow and increases operational complexity.
Using a custom container is a one-time setup that provides maximum reuse with minimal ongoing effort, making it the least implementation effort option in the long run.
Therefore, Option B is the correct and AWS-recommended answer.

NEW QUESTION: 32
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 Serverless Inference for inference. Use SageMaker Inference Recommender for notifications about model quality.
B. Use SageMaker batch transform for inference. Use SageMaker Model Monitor for notifications about model quality.
C. Use SageMaker real-time inference for inference. Use SageMaker Model Monitor for notifications about model quality.
D. Keep using SageMaker Asynchronous Inference for inference. Use SageMaker Inference Recommender for notifications about model quality.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 33
A company has significantly increased the amount of data that is stored as .csv files in an Amazon S3 bucket.
Data transformation scripts and queries are now taking much longer than they used to take.
An ML engineer must implement a solution to optimize the data for query performance.
Which solution will meet this requirement with the LEAST operational overhead?
A. Configure an AWS Lambda function to split the .csv files into smaller objects in the S3 bucket.
B. Configure an AWS Glue job to drop columns that have string type values and to save the results to the S3 bucket.
C. Configure an AWS Glue extract, transform, and load (ETL) job to convert the .csv files to Apache Parquet format.
D. Configure an Amazon EMR cluster to process the data that is in the S3 bucket.
Answer: C (LEAVE A REPLY)
AWS documentation strongly recommends using columnar storage formats to optimize analytical query performance on large datasets stored in Amazon S3. Apache Parquet is a columnar, compressed, and splittable file format that significantly improves query speed and reduces I/O compared to row-based formats such as CSV.
By using AWS Glue to convert CSV files into Parquet format, the company can achieve faster query execution with minimal operational overhead. Glue is fully managed, serverless, and integrates natively with S3, Amazon Athena, and Amazon Redshift Spectrum.
Option A does not improve query efficiency; splitting files still leaves the data in an inefficient row-based format. Option B may reduce data size but does not address the fundamental inefficiency of CSV for analytics. Option D introduces significant operational overhead because Amazon EMR requires cluster provisioning, scaling, and maintenance.
Therefore, converting CSV files to Apache Parquet using AWS Glue ETL is the most efficient and low- maintenance solution.

NEW QUESTION: 34

An ML engineer is designing an AI-powered traffic management system. The system must use near real-time inference to predict congestion and prevent collisions.
The system must also use batch processing to perform historical analysis of predictions over several hours to improve the model. The inference endpoints must scale automatically to meet demand.
Which combination of solutions will meet these requirements? (Select TWO.)
A. Use Amazon SageMaker real-time inference endpoints with automatic scaling based on ConcurrentInvocationsPerInstance.
B. Use AWS Lambda with reserved concurrency and SnapStart to connect to SageMaker endpoints.
C. Use an Amazon SageMaker Processing job for batch historical analysis. Schedule the job with Amazon EventBridge.
D. Use Amazon EC2 Auto Scaling to host containers for batch analysis.
E. Use AWS Lambda for historical analysis.
Answer: (SHOW ANSWER)
For near real-time predictions, AWS documentation recommends Amazon SageMaker real-time inference endpoints. These endpoints support automatic scaling based on metrics such as ConcurrentInvocationsPerInstance, ensuring low latency and high availability during traffic spikes.
For long-running historical analysis, SageMaker Processing jobs are the appropriate solution. Processing jobs are designed for batch workloads, can run for hours, and integrate cleanly with SageMaker pipelines.
Scheduling them with Amazon EventBridge provides a fully managed, scalable, and serverless solution.
AWS Lambda is unsuitable for multi-hour workloads. EC2 Auto Scaling adds unnecessary infrastructure management overhead.
Therefore, Options A and C together meet all requirements and align with AWS best practices.

NEW QUESTION: 35
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 (LEAVE A REPLY)
* 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 a class imbalance in the dataset and interdependencies among features that need to be addressed.
* The solution requires data aggregation from diverse sources for centralized processing.
* Why AWS Lake Formation?
* AWS Lake Formation is 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.
AWS Lake Formation Documentation
AWS Glue for Data Preparation

NEW QUESTION: 36
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. Random search
B. Hyperband
C. Bayesian optimization
D. Grid search
Answer: B (LEAVE A REPLY)

NEW QUESTION: 37
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 B in the same Region.
Which solution will meet this requirement with the LEAST development effort?
A. Use AWS DataSync to replicate the model from Account A 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 Amazon S3 to make a copy of the model. Transfer the copy to Account B.
D. Create an AWS Site-to-Site VPN connection between Account A and Account B to transfer the model.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 38
An insurance company needs to automate claim compliance reviews because human reviews are expensive and error-prone. The company has a large set of claims and a compliance label for each. Each claim consists of a few sentences in English, many of which contain complex related information. Management would like to use Amazon SageMaker built-in algorithms to design a machine learning supervised model that can be trained to read each claim and predict if the claim is compliant or not.
Which approach should be used to extract features from the claims to be used as inputs for the downstream supervised task?

A. Derive a dictionary of tokens from claims in the entire dataset. Apply one-hot encoding to tokens found in each claim of the training set. Send the derived features space as inputs to an Amazon SageMaker built- in supervised learning algorithm.
B. Apply Amazon SageMaker BlazingText in Word2Vec mode to claims in the training set. Send the derived features space as inputs for the downstream supervised task.
C. Apply Amazon SageMaker BlazingText in classification mode to labeled claims in the training set to derive features for the claims that correspond to the compliant and non-compliant labels, respectively.
D. Apply Amazon SageMaker Object2Vec to claims in the training set. Send the derived features space as inputs for the downstream supervised task.
Answer: D (LEAVE A REPLY)
Amazon SageMaker Object2Vec generalizes the Word2Vec embedding technique for words to more complex objects, such as sentences and paragraphs. Since the supervised learning task is at the level of whole claims, for which there are labels, and no labels are available at the word level, Object2Vec needs be used instead of Word2Vec.

NEW QUESTION: 39
A company has significantly increased the amount of data stored as .csv files in an Amazon S3 bucket. Data transformation scripts and queries are now taking much longer than before.
An ML engineer must implement a solution to optimize the data for query performance with the LEAST operational overhead.
Which solution will meet this requirement?
A. Configure an AWS Lambda function to split the .csv files into smaller objects.
B. Configure an AWS Glue job to drop string-type columns and save the results to S3.
C. Configure an AWS Glue ETL job to convert the .csv files to Apache Parquet format.
D. Configure an Amazon EMR cluster to process the data in S3.
Answer: C (LEAVE A REPLY)
AWS strongly recommends converting large CSV datasets into columnar formats such as Apache Parquet to improve query performance. Parquet reduces I/O by reading only the required columns and applies compression, which significantly speeds up analytics workloads.
AWS Glue ETL jobs provide a fully managed, serverless way to perform this conversion with minimal operational overhead. Once converted, the Parquet files can be queried efficiently by services such as Amazon Athena, Redshift Spectrum, and SageMaker processing jobs.
Splitting CSV files does not address inefficient storage format. Dropping columns risks data loss. Amazon EMR introduces infrastructure management overhead and is unnecessary for a straightforward format conversion.
AWS documentation clearly identifies CSV-to-Parquet conversion using Glue ETL as a best practice for scalable analytics.
Therefore, Option C is the correct answer.

NEW QUESTION: 40
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 (LEAVE A REPLY)

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.
Analyze and Visualize - Amazon SageMaker
Transform Data - Amazon SageMaker

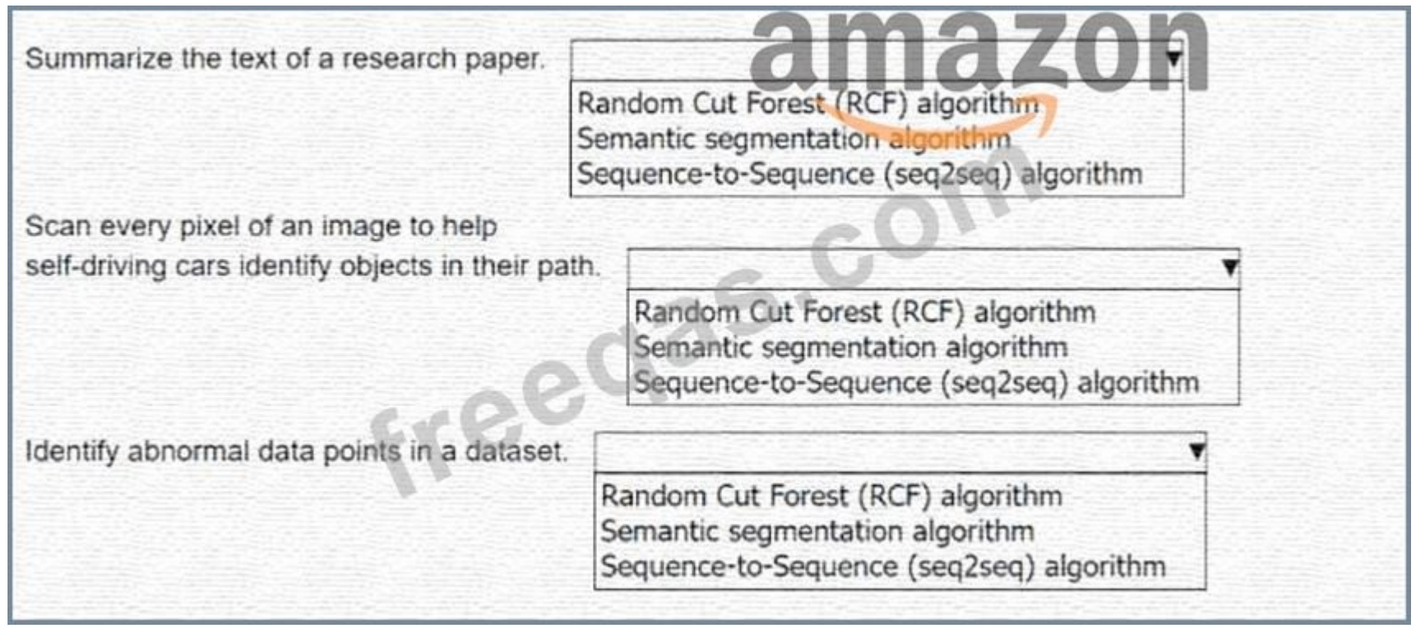
NEW QUESTION: 41
A company receives daily .csv files about customer interactions with its ML model. The company stores the files in Amazon S3 and uses the files to retrain the model. An ML engineer needs to implement a solution to mask credit card numbers in the files before the model is retrained.
Which solution will meet this requirement with the LEAST development effort?
A. Create a discovery job in Amazon Macie. Configure the job to find and mask sensitive data.
B. Create Apache Spark code to run on an AWS Glue job. Use the Sensitive Data Detection functionality in AWS Glue to find and mask sensitive data.
C. Create Apache Spark code to run on an Amazon EC2 instance. Program the code to perform an operation to find and mask sensitive data.
D. Create Apache Spark code to run on an AWS Glue job. Program the code to perform a regex operation to find and mask sensitive data.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 42
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 (LEAVE A REPLY)
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.

NEW QUESTION: 43
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 (LEAVE A REPLY)
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.

NEW QUESTION: 44
An ML engineer must choose the appropriate Amazon SageMaker algorithm to solve specific AI problems.
Select the correct SageMaker built-in algorithm from the following list for each use case. Each algorithm should be selected one time.
* Random Cut Forest (RCF) algorithm
* Semantic segmentation algorithm
* Sequence-to-Sequence (seq2seq) algorithm



Answer:

Explanation:

Use case 1:

Summarize the text of a research paper

Sequence-to-Sequence (seq2seq) algorithm

Why:

Seq2seq models are designed for natural language generation tasks such as text summarization, translation, and paraphrasing. AWS documentation explicitly lists text summarization as a primary use case for the SageMaker seq2seq algorithm.

Use case 2:

Scan every pixel of an image to help self-driving cars identify objects in their path
Semantic segmentation algorithm
Why:
Semantic segmentation performs pixel-level classification, assigning a class label to every pixel in an image.
This is exactly what is required for applications such as autonomous driving, road scene understanding, and object boundary detection.
Use case 3:
Identify abnormal data points in a dataset
Random Cut Forest (RCF) algorithm
Why:
Random Cut Forest is an unsupervised anomaly detection algorithm. AWS SageMaker RCF is purpose-built to identify outliers, unusual patterns, and anomalies in numerical datasets, making it ideal for fraud detection, monitoring, and abnormal data point detection.

NEW QUESTION: 45
An ML engineer needs to encrypt all data in transit when an ML training job runs. The ML engineer must ensure that encryption in transit is applied to processes that Amazon SageMaker uses during the training job.
Which solution will meet these requirements?
A. Encrypt communication between nodes for batch processing.
B. Specify an AWS Key Management Service (AWS KMS) key during creation of the training job request.
C. Encrypt communication between nodes in a training cluster.
D. Specify an AWS Key Management Service (AWS KMS) key during creation of the SageMaker domain.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 46
A machine learning team has several large CSV datasets in Amazon S3. Historically, models built with the Amazon SageMaker Linear Learner algorithm have taken hours to train on similar-sized datasets. The team's leaders need to accelerate the training process.
What can a machine learning specialist do to address this concern?
A. Use Amazon SageMaker Pipe mode.
B. Use Amazon Machine Learning to train the models.
C. Use Amazon Kinesis to stream the data to Amazon SageMaker.
D. Use AWS Glue to transform the CSV dataset to the JSON format.
Answer: A (LEAVE A REPLY)
Amazon SageMaker Pipe mode streams the data directly to the container, which improves the performance of training jobs. In Pipe mode, your training job streams data directly from Amazon S3. Streaming can provide faster start times for training jobs and better throughput. With Pipe mode, you also reduce the size of the Amazon EBS volumes for your training instances.

NEW QUESTION: 47

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. 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.
B. 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.
C. 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.
D. 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.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 48
A gaming company needs to deploy a natural language processing (NLP) model to moderate a chat forum in a game. The workload experiences heavy usage during evenings and weekends but minimal activity during other hours.
Which solution will meet these requirements MOST cost-effectively?
A. Use an Amazon SageMaker AI batch transform job with fixed capacity.
B. Use Amazon SageMaker Serverless Inference.
C. Use a single Amazon EC2 GPU instance with reserved capacity.
D. Use Amazon SageMaker Asynchronous Inference.
Answer: (SHOW ANSWER)
The key requirements in this scenario are variable traffic patterns and cost efficiency. The workload has unpredictable spikes during evenings and weekends, followed by long periods of low or no usage. According to AWS Machine Learning documentation, Amazon SageMaker Serverless Inference is specifically designed for such use cases.
SageMaker Serverless Inference automatically provisions, scales, and shuts down compute resources based on incoming inference requests. Customers are billed only for the compute time used during inference, not for idle resources. This makes it highly cost-effective for workloads with intermittent or spiky traffic, such as real- time chat moderation in gaming environments.
Option A is incorrect because batch transform jobs are intended for offline, large-scale inference and require fixed capacity during job execution. They are not suitable for real-time NLP moderation.
Option C is also incorrect because reserving an EC2 GPU instance incurs continuous costs regardless of utilization. This would be inefficient given the long idle periods described in the scenario.
Option D, SageMaker Asynchronous Inference, is designed for workloads with long processing times or large payloads and still requires endpoint provisioning. While it can handle traffic spikes, it does not scale down to zero in the same cost-efficient manner as Serverless Inference.
Therefore, Amazon SageMaker Serverless Inference is the most cost-effective and operationally efficient solution for deploying an NLP moderation model with highly variable usage patterns.

NEW QUESTION: 49
A company regularly receives new training data from a 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 AI pipeline to retrain the model. An ML engineer needs to run the pipeline automatically 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 AI training instance and initiate training.
B. Create an AWS Lambda function that scans the S3 bucket and initiates the pipeline when new data is uploaded.
C. Create an Amazon EventBridge rule that matches S3 upload events and configures the SageMaker pipeline as the target.
D. Use Amazon Managed Workflows for Apache Airflow (MWAA) to orchestrate the pipeline when new data is uploaded.
Answer: (SHOW ANSWER)

AWS best practices recommend event-driven architectures to automate ML workflows with minimal operational overhead. Amazon EventBridge natively integrates with Amazon S3 and Amazon SageMaker Pipelines, making it the most efficient solution for triggering retraining when new data arrives.
Amazon S3 automatically emits object creation events. By creating an EventBridge rule that listens for these events and targets a SageMaker Pipeline execution, the pipeline can start immediately when new training data is uploaded. This solution requires no custom code, no polling, and no infrastructure management.
Option A is incorrect because S3 lifecycle rules manage storage transitions, not workflow execution. Option B introduces custom code and periodic scanning, which increases operational complexity and cost. Option D (MWAA) is powerful but requires maintaining an Airflow environment and is unnecessary for a simple event- based trigger.
AWS documentation explicitly highlights EventBridge + SageMaker Pipelines as the recommended pattern for automated retraining workflows triggered by data arrival.
Therefore, Option C is the correct and AWS-verified answer.

NEW QUESTION: 50
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 (LEAVE A REPLY)
Predicting apartment prices is a regression problem, where the target variable is continuous. AWS documentation states that classification metrics such as accuracy, AUC, and F1 score are not appropriate for regression tasks.
Mean Absolute Error (MAE) measures the average absolute difference between predicted values and actual values. MAE is easy to interpret because it is expressed in the same units as the target variable (for example, dollars), making it especially useful for business-facing problems like price prediction.
AWS best practices recommend MAE for evaluating regression models when understanding average prediction error magnitude is important and when robustness to outliers is desired.
Therefore, Option D is the correct and AWS-aligned answer.

NEW QUESTION: 51
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 (LEAVE A REPLY)
An oscillating loss pattern during training with stochastic gradient descent (SGD) is a strong indicator that the learning rate is too high. When the learning rate is excessive, the optimizer takes overly large steps during gradient updates, causing the model to repeatedly overshoot the optimal minimum of the loss function. This results in unstable convergence behavior, where training and validation loss decrease briefly and then increase again in a repeating cycle.
AWS Machine Learning documentation and general deep learning best practices recommend reducing the learning rate when training loss and validation loss both remain high and fluctuate rather than steadily decreasing. Lowering the learning rate allows the optimizer to take smaller, more precise steps toward the minimum, leading to smoother convergence and improved generalization on the test dataset.
Option A, early stopping, is used primarily to prevent overfitting when validation loss increases while training loss continues to decrease. In this scenario, both losses remain high and unstable, indicating an optimization issue rather than overfitting.

Option B is incorrect because increasing the test set size does not affect the training dynamics or convergence behavior of the model.
Option C would worsen the problem, as increasing the learning rate would further amplify oscillations and instability.
Therefore, decreasing the learning rate is the correct corrective action to stabilize SGD training and improve model performance.

NEW QUESTION: 52
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 (LEAVE A REPLY)
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.

NEW QUESTION: 53
An ML engineer is collecting data to train a classification ML model by using Amazon SageMaker AI. The target column can have two possible values: Class A or Class B. The ML engineer wants to ensure that the number of samples for both Class A and Class B are balanced, without losing any existing training data. The ML engineer must test the balance of the training data.
Which solution will meet this requirement?
A. Use SageMaker Clarify to check for class imbalance (CI). If the value is equal to 0, then use random undersampling in SageMaker Data Wrangler to balance the classes.
B. Use SageMaker Clarify to check for class imbalance (CI). If the value is greater than 0, then use synthetic minority oversampling technique (SMOTE) in SageMaker Data Wrangler to balance the classes.
C. Use SageMaker JumpStart to generate a class imbalance (CI) report. If the value is greater than 0, then use random undersampling in SageMaker Studio to balance the classes.
D. Use SageMaker JumpStart to generate a class imbalance (CI) report. If the value is equal to 0, then use synthetic minority oversampling technique (SMOTE) in SageMaker Studio to balance the classes.
Answer: B (LEAVE A REPLY)
The requirement has two key constraints: detect class imbalance and balance classes without losing any existing data. AWS provides Amazon SageMaker Clarify as the native tool to detect pre-training bias, including class imbalance (CI). CI measures differences in label distributions between classes, and a CI value greater than 0 indicates imbalance.
Once imbalance is detected, the engineer must rebalance the dataset without discarding data. Random undersampling would remove samples from the majority class, violating the requirement.
Instead, oversampling is required. SMOTE (Synthetic Minority Oversampling Technique) creates synthetic samples for the minority class, preserving all original data while improving class balance.
Amazon SageMaker Data Wrangler natively supports SMOTE, making it the correct AWS-managed tool for this preprocessing task.
Options C and D are incorrect because SageMaker JumpStart is used for pretrained models and solutions, not for bias detection reporting. Option A is incorrect because it uses undersampling and misinterprets CI = 0 (which actually indicates no imbalance).
Therefore, detecting imbalance with SageMaker Clarify and correcting it using SMOTE in Data Wrangler is the correct solution.

NEW QUESTION: 54
A company ingests sales transaction data using Amazon Data Firehose into Amazon OpenSearch Service. The Firehose buffer interval is set to 60 seconds.
The company needs sub-second latency for a real-time OpenSearch dashboard.
Which architectural change will meet this requirement?
A. Use zero buffering in the Firehose stream and tune the PutRecordBatch batch size.

B. Replace Firehose with AWS DataSync and enhanced fan-out consumers.
C. Increase the Firehose buffer interval to 120 seconds.
D. Replace Firehose with Amazon SQS.
Answer: A (LEAVE A REPLY)
Amazon Data Firehose supports near real-time delivery by configuring the buffer interval and buffer size.
AWS documentation states that setting the buffer interval to the minimum (as low as 1 second) enables low- latency ingestion.
By using zero or minimal buffering and tuning PutRecordBatch, data is delivered to OpenSearch almost immediately, enabling sub-second dashboard updates.
DataSync and SQS are not designed for real-time streaming analytics. Increasing the buffer interval worsens latency.
AWS explicitly recommends Firehose with minimal buffering for real-time OpenSearch ingestion.
Therefore, Option A is the correct and AWS-verified solution.

NEW QUESTION: 55
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 (LEAVE A REPLY)
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 offers Managed 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 using Managed Warm Pools, the company can significantly reduce startup latency for consecutive training jobs, ensuring faster experimentation cycles with minimal operational overhead.
AWS Documentation: Managed Warm Pools
AWS Blog: Reduce ML Model Training Job Startup Time

NEW QUESTION: 56
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: (SHOW ANSWER)
The primary requirement in this scenario is achieving sub-second latency for a real-time analytics dashboard powered by Amazon OpenSearch Service. The current architecture uses Amazon Data Firehose, which buffers incoming records based on time or size before delivering them to the destination. A buffer interval of
60 seconds introduces unavoidable latency, making it unsuitable for near-real-time or sub-second use cases.
According to AWS documentation, reducing or eliminating buffering in Firehose is the correct approach when low-latency ingestion is required. Setting the Firehose buffer interval to zero seconds forces Firehose to deliver records as soon as they are received. Additionally, tuning the PutRecordBatch batch size allows efficient ingestion while minimizing delivery delay. This configuration is explicitly recommended for latency- sensitive analytics pipelines.
Option B is incorrect because AWS DataSync is designed for batch-oriented data transfers between storage systems, not real-time streaming. Enhanced fan-out consumers are a feature of Amazon Kinesis Data Streams, not DataSync, making this option invalid.
Option C directly contradicts the requirement. Increasing the buffer interval from 60 seconds to 120 seconds would further increase latency and degrade real-time performance.
Option D is also incorrect because Amazon SQS is a message queueing service, not a streaming ingestion service optimized for indexing data into OpenSearch with minimal latency. Using SQS would add additional processing layers and would not inherently provide sub-second ingestion into OpenSearch.
Therefore, using zero buffering in the Firehose stream and tuning the PutRecordBatch batch size is the only change that aligns with AWS best practices for achieving sub-second latency in real-time analytics pipelines.

NEW QUESTION: 57
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: (SHOW ANSWER)
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.

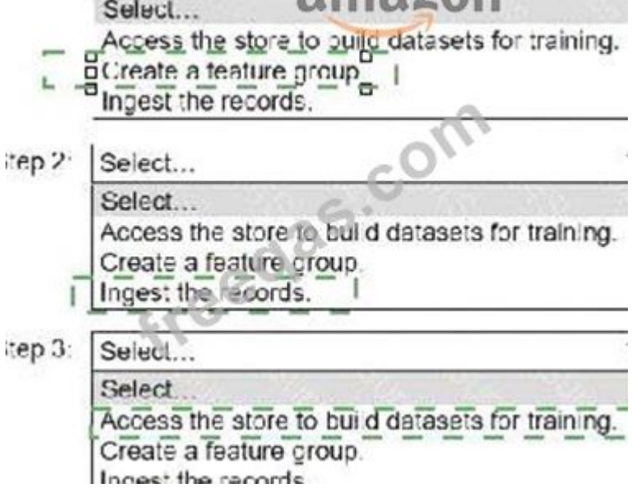
NEW QUESTION: 58

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.

NEW QUESTION: 59

A company wants to use large language models (LLMs) supported by Amazon Bedrock to develop a chat interface for internal technical documentation.
The documentation consists of dozens of text files totaling several megabytes and is updated frequently.
Which solution will meet these requirements MOST cost-effectively?

A. Train a new LLM in Amazon Bedrock using the documentation.
B. Use Amazon Bedrock guardrails to integrate documentation.
C. Fine-tune an LLM in Amazon Bedrock with the documentation.
D. Upload the documentation to an Amazon Bedrock knowledge base and use it as context during inference.

Answer: D (LEAVE A REPLY)

AWS recommends Retrieval Augmented Generation (RAG) using Amazon Bedrock knowledge bases as the most cost-effective solution for incorporating frequently updated documents into LLM-powered applications.

A Bedrock knowledge base allows the company to store documents in Amazon S3, index them using vector embeddings, and retrieve relevant context dynamically at inference time. This approach eliminates the need for retraining or fine-tuning the model when documents change.
Training or fine-tuning an LLM is expensive, time-consuming, and unnecessary for frequently changing data.
Bedrock guardrails are designed for safety and policy enforcement, not knowledge integration.
AWS documentation explicitly states that knowledge bases are the preferred method for dynamic, updatable enterprise content in chat applications.
Therefore, Option D is the correct and most cost-effective solution.

NEW QUESTION: 60
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 (LEAVE A REPLY)
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

NEW QUESTION: 61
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 (LEAVE A REPLY)
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.

NEW QUESTION: 62
An IoT company uses Amazon SageMaker to train and test an XGBoost model for object detection. ML engineers need to monitor performance metrics when they train the model with variants in hyperparameters. The ML engineers also need to send Short Message Service (SMS) text messages after training is complete.
Which solution will meet these requirements?
A. Use AWS CloudTrail to monitor performance metrics. Use Amazon Simple Notification Service (Amazon SNS) for message delivery.
B. Use Amazon CloudWatch to monitor performance metrics. Use Amazon Simple Queue Service (Amazon SQS) for message delivery.
C. Use Amazon CloudWatch to monitor performance metrics. Use Amazon Simple Notification Service (Amazon SNS) for message delivery.
D. Use AWS CloudTrail to monitor performance metrics. Use Amazon Simple Queue Service (Amazon SQS) for message delivery.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 63
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 Amazon CloudWatch. Configure an Amazon CloudFront rule to send the email message.
B. Log the metrics from the Lambda function to AWS CloudTrail. Configure a CloudTrail trail to send the email message.
C. Log the metrics from the Lambda function to Amazon CloudFront. Configure an Amazon CloudWatch alarm to send the email message.
D. Log the metrics from the Lambda function to Amazon CloudWatch. Configure a CloudWatch alarm to send the email message.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 64

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. Investigate and reduce the sources of model bias.
B. Increase the number of layers.
C. Enable early stopping on the model.
D. Increase dropout in the layers.
E. Increase the number of neurons.
Answer: (SHOW ANSWER)

NEW QUESTION: 65
A company wants to provide services to help other businesses label images. The company wants its labeling specialists to complete human labeling tasks on AWS. How should the company register the labeling specialists to receive tasks on AWS?
A. Use AWS Data Exchange.
B. Create and use an internal workforce in Amazon SageMaker Ground Truth.
C. Create and use Amazon Mechanical Turk entities in an Amazon SageMaker human loop.
D. Use the Amazon Mechanical Turk website.
Answer: (SHOW ANSWER)
To enable labeling specialists within the company to perform tasks, the correct solution is to create and use an internal workforce in Amazon SageMaker Ground Truth. This allows the company to securely register and manage its own labeling team to receive and complete human labeling tasks.

NEW QUESTION: 66
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 catalog the models.
D. Use the SageMaker Model Registry and unique tags for each model version.
Answer: (SHOW ANSWER)
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 into Model 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.
AWS Documentation: SageMaker Model Registry
AWS Blog: Model Registry Features and Usage

NEW QUESTION: 67
A company has an ML model that uses historical transaction data to predict customer behavior.
An ML engineer is optimizing the model in Amazon SageMaker to enhance the model's predictive accuracy. The ML engineer must examine the input data and the resulting predictions to identify trends that could skew the model's performance across different demographics.
Which solution will provide this level of analysis?
A. Create AWS Glue DataBrew recipes to correct the data based on statistics from the model output.
B. Use SageMaker Clarify to evaluate the model and training data for underlying patterns that might affect accuracy.
C. Use Amazon CloudWatch to monitor network metrics and CPU metrics for resource optimization during model training.
D. Create AWS Lambda functions to automate data pre-processing and to ensure consistent quality of input data for the model.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 68
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 SageMaker batch transform. Use SageMaker Model Monitor to monitor the data quality and to send alerts.
D. 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.
Answer: (SHOW ANSWER)

NEW QUESTION: 69
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 blue/green deployments with all-at-once traffic shifting.
B. Set up SageMaker Debugger and create a custom rule.
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 (LEAVE A REPLY)

NEW QUESTION: 70
A company is using Amazon SageMaker AI to build an ML model to predict customer behavior. The company needs to explain the bias in the model to an auditor. The explanation must focus on demographic data of the customers.
Which solution will meet these requirements?
A. Use SageMaker Clarify to generate a bias report. Send the report to the auditor.
B. Use AWS Glue DataBrew to create a job to detect drift in the model's data quality. Send the job output to the auditor.
C. Use Amazon QuickSight integration with SageMaker AI to generate a bias report. Send the report to the auditor.
D. Use Amazon CloudWatch metrics from the SageMaker AI namespace to create a bias dashboard. Share the dashboard with the auditor.
Answer: (SHOW ANSWER)
AWS documentation identifies Amazon SageMaker Clarify as the primary service for detecting, measuring, and explaining bias in ML models, particularly across demographic and sensitive attributes such as age, gender, and location. Clarify can analyze bias before training, after training, and during inference, making it suitable for audit and compliance requirements.
SageMaker Clarify generates bias reports using established fairness metrics such as difference in positive proportions, disparate impact, and conditional demographic disparity. These reports are exportable and auditor-friendly, directly meeting the requirement to explain bias to an external party.
AWS Glue DataBrew focuses on data preparation and quality, not bias detection. Amazon QuickSight does not provide ML fairness metrics. Amazon CloudWatch captures operational metrics, not demographic bias indicators.
AWS best practices explicitly recommend SageMaker Clarify for model transparency, fairness evaluation, and regulatory reporting.
Therefore, Option A is the correct and AWS-verified solution.

NEW QUESTION: 71
A company uses Amazon SageMaker AI to create ML models. The data scientists need fine-grained control of ML workflows, DAG visualization, experiment history, and model governance for auditing and compliance.
Which solution will meet these requirements?
A. Use AWS CodePipeline with SageMaker Studio and SageMaker ML Lineage Tracking.
B. Use AWS CodePipeline with SageMaker Experiments.
C. Use SageMaker Pipelines with SageMaker Studio and SageMaker ML Lineage Tracking.
D. Use SageMaker Pipelines with SageMaker Experiments.
Answer: C (LEAVE A REPLY)
Amazon SageMaker Pipelines provides native orchestration of ML workflows with fine-grained control, DAG-based visualization, and seamless integration with SageMaker Studio. AWS documentation explicitly states that Pipelines is designed for end-to-end ML workflow automation and visualization.
SageMaker ML Lineage Tracking records relationships between datasets, models, training jobs, and endpoints, enabling full auditability and governance, which is essential for compliance.
SageMaker Experiments tracks experiment metrics but does not provide lineage-level governance.
CodePipeline is a general CI/CD service and lacks ML-specific DAG visualization and lineage tracking.
AWS best practices recommend combining SageMaker Pipelines + SageMaker Studio + ML Lineage Tracking for enterprise-grade ML workflow management.
Therefore, Option C is the correct and AWS-verified solution.

NEW QUESTION: 72
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 (LEAVE A REPLY)
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.

NEW QUESTION: 73
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. Create a materialized view with masking logic on top of the database. Grant the necessary read permissions to the data scientist.
B. 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.
C. Unload the Amazon Redshift data to Amazon S3. Create an AWS Glue job to anonymize the data. Share the dataset with the data scientist.
D. Configure dynamic data masking policies to control how sensitive data is shared with the data scientist at query time.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 74
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: (SHOW ANSWER)
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.

NEW QUESTION: 75
A company uses AWS CodePipeline to orchestrate a continuous integration and continuous delivery (CI/CD) pipeline for ML models and applications.
Select and order the steps from the following list to describe a CI/CD process for a successful deployment.
Select each step one time. (Select and order FIVE.)
. CodePipeline deploys ML models and applications to production.
CodePipeline detects code changes and starts to build automatically.
. Human approval is provided after testing is successful.

. The company builds and deploys ML models and applications to staging servers for testing.
. The company commits code changes or new training datasets to a Git repository.

Step 1: Select...

- Select...
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

Step 2: Select...

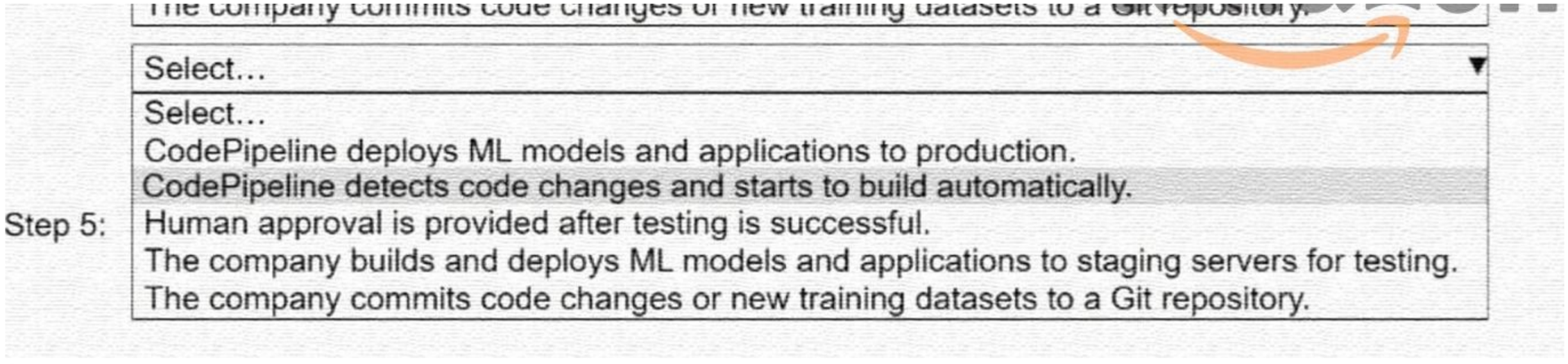
- Select...
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

Step 3: Select...

- Select...
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

Step 4: Select...

- Select...
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.


The company commits code changes of new training datasets to a Git repository.
Select...
Step 5:
Select...
CodePipeline deploys ML models and applications to production.
CodePipeline detects code changes and starts to build automatically.
Human approval is provided after testing is successful.
The company builds and deploys ML models and applications to staging servers for testing.
The company commits code changes or new training datasets to a Git repository.

Answer:

Step 1: Select…

- Select…
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

Step 2: Select…

- Select…
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

Step 3: Select…

- Select…
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

Step 4: Select…

- Select…
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

Step 5: Select…

- Select…
- CodePipeline deploys ML models and applications to production.
- CodePipeline detects code changes and starts to build automatically.
- Human approval is provided after testing is successful.
- The company builds and deploys ML models and applications to staging servers for testing.
- The company commits code changes or new training datasets to a Git repository.

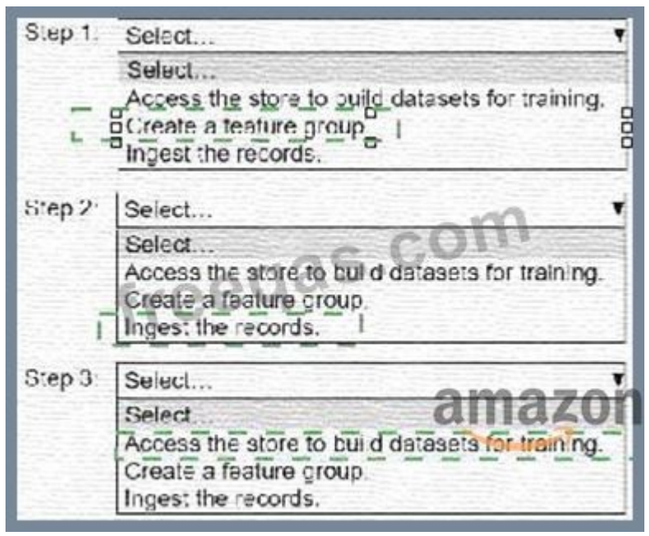
Explanation:
Step 1:

The company commits code changes or new training datasets to a Git repository.
This is the trigger point. A source code or data change initiates the CI/CD pipeline.
Step 2:
CodePipeline detects code changes and starts to build automatically.
CodePipeline monitors the Git repository (for example, AWS CodeCommit, GitHub, or Bitbucket) and automatically triggers the pipeline when changes are detected.
Step 3:
The company builds and deploys ML models and applications to staging servers for testing.
The pipeline runs build, training, and test stages (often using AWS CodeBuild and SageMaker) and deploys artifacts to a staging or test environment for validation.
Step 4:
Human approval is provided after testing is successful.
A manual approval action is a best practice for ML workflows to ensure governance, compliance, and quality checks before production deployment.
Step 5:
CodePipeline deploys ML models and applications to production.
After approval, the pipeline automatically deploys the validated model or application to the production environment.

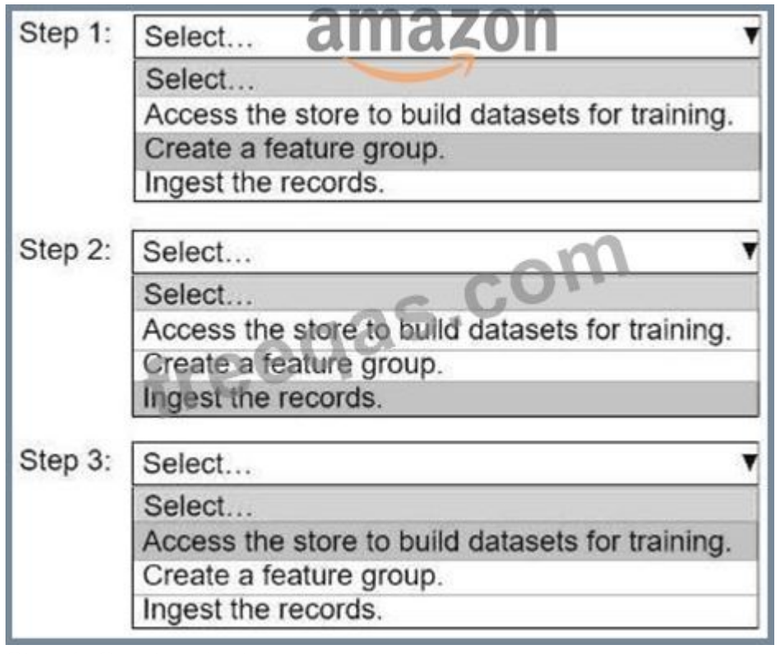
NEW QUESTION: 76
An ML engineer needs to use AWS CloudFormation to create an ML model that an Amazon SageMaker endpoint will host.
Which resource should the ML engineer declare in the CloudFormation template to meet this requirement?
A. AWS::SageMaker::Model
B. AWS::SageMaker::NotebookInstance
C. AWS::SageMaker::Pipeline
D. AWS::SageMaker::Endpoint
Answer: A (LEAVE A REPLY)

NEW QUESTION: 77
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.

NEW QUESTION: 78

A data scientist is working on optimizing a model during the training process by varying multiple parameters. The data scientist observes that, during multiple runs with identical parameters, the loss function converges to different, yet stable, values.

What should the data scientist do to improve the training process?

A. Increase the learning rate. Keep the batch size the same.

B. Decrease the learning rate. Reduce the batch size.
C. Decrease the learning rate. Keep the batch size the same.
D. Do not change the learning rate. Increase the batch size.
Answer: (SHOW ANSWER)
It is most likely that the loss function is very curvy and has multiple local minima where the training is getting stuck. Decreasing the batch size would help the data scientist stochastically get out of the local minima saddles. Decreasing the learning rate would prevent overshooting the global loss function minimum.

NEW QUESTION: 79
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. 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.
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. 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.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 80
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 (LEAVE A REPLY)
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.

NEW QUESTION: 81
A company is building an Amazon SageMaker AI pipeline for an ML model. The pipeline uses distributed processing and training.
An ML engineer needs to encrypt network communication between instances that run distributed jobs. The ML engineer configures the distributed jobs to run in a private VPC.
What should the ML engineer do to meet the encryption requirement?
A. Enable network isolation.
B. Configure traffic encryption by using security groups.
C. Enable inter-container traffic encryption.
D. Enable VPC flow logs.
Answer: C (LEAVE A REPLY)

In distributed training and processing jobs, multiple instances and containers communicate with each other over the network to exchange gradients, parameters, and intermediate results. Even when jobs run inside a private VPC, network traffic between instances is not automatically encrypted at the application layer.
AWS documentation for Amazon SageMaker specifies that inter-container traffic encryption is the supported mechanism for encrypting data in transit between containers that participate in distributed training or processing jobs. When enabled, SageMaker uses TLS to encrypt all communication between containers across instances, ensuring confidentiality and compliance with security requirements.
Option A (network isolation) prevents containers from making outbound network calls but does not encrypt traffic between distributed instances. Option B is incorrect because security groups control traffic access, not encryption. Option D (VPC flow logs) is a monitoring feature and does not provide encryption.
Therefore, enabling inter-container traffic encryption is the correct and AWS-recommended solution.

NEW QUESTION: 82
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. 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.
B. 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.
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: A (LEAVE A REPLY)
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.

NEW QUESTION: 83
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. Mean absolute error (MAE)
B. Accuracy
C. Area Under the ROC Curve (AUC)
D. F1 score
Answer: A (LEAVE A REPLY)

NEW QUESTION: 84
An ML engineer is setting up an Amazon SageMaker AI pipeline for an ML model. The pipeline must automatically initiate a re-training job if any data drift is detected.
How should the ML engineer set up the pipeline to meet this requirement?
A. Use an AWS Glue crawler and an AWS Glue extract, transform, and load (ETL) job to detect data drift.
Use AWS Glue triggers to automate the retraining job.
B. Use Amazon Managed Service for Apache Flink to detect data drift. Use an AWS Lambda function to automate the re-training job.
C. Use SageMaker Model Monitor to detect data drift. Use an AWS Lambda function to automate the re- training job.
D. Use Amazon Quick Suite (previously known as Amazon QuickSight) anomaly detection to detect data drift. Use an AWS Step Functions workflow to automate the re-training job.
Answer: (SHOW ANSWER)
AWS provides Amazon SageMaker Model Monitor as a native solution for detecting data drift and model quality issues in production ML pipelines. Model Monitor continuously analyzes incoming

inference data and compares it with baseline training data to identify schema drift, feature distribution drift, and data quality anomalies.
When drift thresholds are violated, Model Monitor generates CloudWatch metrics and alerts. These alerts can directly trigger an AWS Lambda function, which can then programmatically initiate a SageMaker retraining job or start a SageMaker Pipeline execution. This design is explicitly documented by AWS as the recommended architecture for automated retraining workflows.
Option A is incorrect because AWS Glue is a data integration service and does not provide ML-specific drift detection capabilities.
Option B is incorrect because Apache Flink is designed for stream processing, not ML data drift detection.
Option D is incorrect because Amazon QuickSight anomaly detection is intended for business intelligence metrics, not ML feature drift.
Therefore, using SageMaker Model Monitor with AWS Lambda automation is the correct, AWS-native solution for drift-driven retraining.

NEW QUESTION: 85
A company needs an AWS solution that will automatically create versions of ML models as the models are created.
Which solution will meet this requirement?
A. Amazon SageMaker ML Lineage Tracking
B. Model packages from Amazon SageMaker Marketplace
C. Amazon Elastic Container Registry (Amazon ECR)
D. Amazon SageMaker Model Registry
Answer: (SHOW ANSWER)

NEW QUESTION: 86
An ML engineer notices class imbalance in an image classification training job.
What should the ML engineer do to resolve this issue?
A. Apply random oversampling on the dataset.
B. Transform some of the images in the dataset.
C. Reduce the size of the dataset.
D. Apply random data splitting on the dataset.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 87
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 Canvas user must have permissions to access the S3 bucket where the model artifacts are stored.
B. The model must be registered in the SageMaker Model Registry.
C. The ML engineer must host the model on AWS Marketplace.
D. The ML engineer and the Canvas user must be in separate SageMaker domains.
E. The ML engineer must deploy the model to a SageMaker endpoint.
Answer: A,B (LEAVE A REPLY)

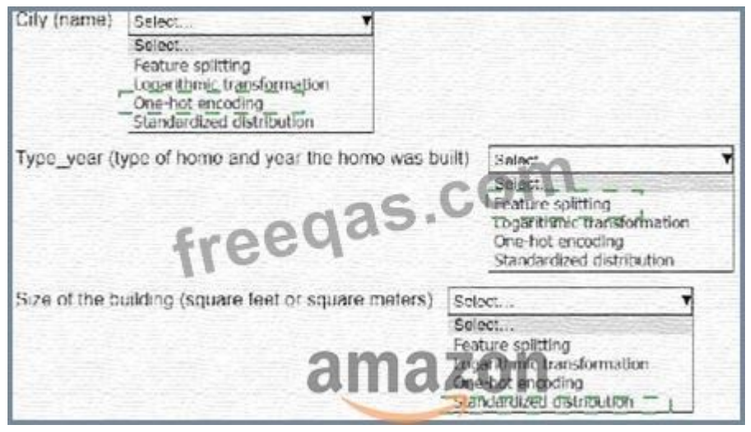
NEW QUESTION: 88
A company stores training data as a .csv file in an Amazon S3 bucket. The company must encrypt the data and must control which applications have access to the encryption key. Which solution will meet these requirements?
A. Create a new SSH access key. Use the AWS Encryption CLI with a reference to the new access key to encrypt the file.

B. Create a new API key by using the Amazon API Gateway CreateApiKey API operation. Use the AWS CLI with a reference to the new API key to encrypt the file.
C. Create a new IAM role. Attach a policy that allows the AWS Key Management Service (AWS KMS) GenerateDataKey action. Use the role to encrypt the file.
D. Create a new AWS Key Management Service (AWS KMS) key. Use the AWS Encryption CLI with a reference to the new KMS key to encrypt the file.
Answer: D (LEAVE A REPLY)
The correct approach is to create a new AWS KMS key and use the AWS Encryption CLI to encrypt the file with that key. This ensures the data in S3 is encrypted, and access to the encryption key can be controlled through KMS key policies and IAM permissions, meeting both encryption and access control requirements.

NEW QUESTION: 89
An ML engineer wants to re-train an XGBoost model at the end of each month. A data team prepares the training data. The training dataset is a few hundred megabytes in size. When the data is ready, the data team stores the data as a new file in an Amazon S3 bucket.
The ML engineer needs a solution to automate this pipeline. The solution must register the new model version in Amazon SageMaker Model Registry within 24 hours.
Which solution will meet these requirements?
A. Create an AWS Lambda function that runs one time each week to poll the S3 bucket for new files.
Invoke the Lambda function asynchronously. Configure the Lambda function to start the pipeline if the function detects new data.
B. Create an Amazon CloudWatch rule that runs on a schedule to start the pipeline every 30 days.
C. Create an S3 Lifecycle rule to start the pipeline every time a new object is uploaded to the S3 bucket.
D. Create an Amazon EventBridge rule to start an AWS Step Functions TrainingStep every time a new object is uploaded to the S3 bucket.
Answer: D (LEAVE A REPLY)
The requirement is event-driven automation when new data arrives in Amazon S3, followed by training and model registration. Amazon EventBridge natively supports S3 object creation events and can trigger downstream workflows immediately.
By using EventBridge to start an AWS Step Functions workflow that includes a training step and a SageMaker Model Registry registration step, the pipeline runs automatically as soon as new data is uploaded-well within the 24-hour requirement.
Option A introduces unnecessary polling and delay. Option B is time-based and does not ensure alignment with data readiness. Option C is invalid because S3 Lifecycle rules manage object transitions, not workflow execution.
Therefore, EventBridge-triggered Step Functions is the correct solution.

NEW QUESTION: 90
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 each unique 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.

NEW QUESTION: 91

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

An ML engineer needs to encrypt network communication between instances that run distributed jobs. The ML engineer configures the distributed jobs to run in a private VPC.
What should the ML engineer do to meet the encryption requirement?
A. Enable network isolation.
B. Configure traffic encryption by using security groups.
C. Enable inter-container traffic encryption.
D. Enable VPC flow logs.
Answer: C (LEAVE A REPLY)
In Amazon SageMaker, distributed training and distributed processing jobs often involve multiple instances exchanging data over the network. By default, when these jobs run inside a VPC, network traffic remains private but is not automatically encrypted between instances. When compliance or security requirements mandate encryption of in-transit data, additional configuration is required.
The correct solution is to enable inter-container traffic encryption, which ensures that all network communication between containers running on different instances is encrypted using TLS. Amazon SageMaker provides a built-in feature for this purpose. When inter-container traffic encryption is enabled, SageMaker automatically configures secure communication channels between all nodes participating in a distributed job, including training clusters and processing jobs.
Option A (Network isolation) is incorrect because network isolation prevents containers from making outbound network calls and accessing the internet. It does not encrypt traffic between instances.
Option B (Security groups) is incorrect because security groups control network access and traffic flow, not encryption. They can restrict which instances can communicate, but they do not provide data-in-transit encryption.
Option D (VPC flow logs) is incorrect because VPC flow logs are used for monitoring and auditing network traffic, not for encrypting it.
AWS documentation explicitly states that enabling inter-container traffic encryption is the recommended and supported approach for encrypting data exchanged between instances during distributed SageMaker jobs. This feature aligns with enterprise security best practices and regulatory requirements for protecting sensitive ML training data in transit.
Therefore, Option C is the only solution that directly fulfills the encryption requirement for distributed SageMaker workloads.

NEW QUESTION: 92
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 AI compute costs reach a specific threshold.
Which solution will meet these requirements?
A. Add resource tagging by editing the SageMaker AI user profile in the SageMaker AI domain. Configure AWS Cost Explorer to send an alert when the threshold is reached.
B. Add resource tagging by editing the SageMaker AI user profile in the SageMaker AI 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 (LEAVE A REPLY)
AWS best practices for cost governance recommend using resource tagging combined with AWS Budgets to track and alert on service-level spending. By adding tags at the SageMaker Studio user profile level, all compute resources launched by users inherit those tags automatically.
AWS Budgets supports threshold-based alerts, unlike AWS Cost Explorer, which is primarily used for historical analysis and visualization. Budgets can trigger notifications via email or Amazon SNS when spending exceeds defined limits.
IAM profiles are unrelated to cost tracking, making options C and D invalid. Therefore, tagging SageMaker user profiles and using AWS Budgets is the correct solution.

NEW QUESTION: 93
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 Pipelines. When a model version is registered, use the AWS SDK to change the approval status to "Approved."
C. Use SageMaker ML Lineage Tracking on the central model registry. Create tracking entities for the approval process.
D. Use SageMaker Model Monitor to evaluate the performance of the model and to manage the approval.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 94
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 Amazon EventBridge to invoke the pipelines and the AWS Glue jobs in the desired order.
C. Use processing steps in SageMaker Pipelines. Configure inputs that point to the Amazon Resource Names (ARNs) of the AWS Glue jobs.
D. Use Callback steps in SageMaker Pipelines to start the AWS Glue workflow and to stop the pipelines until the AWS Glue jobs finish running.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 95
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 AWS Glue to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.
B. Use Amazon Athena to create the data ingestion pipelines. Use an Amazon SageMaker notebook 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 Data Firehose to create the data ingestion pipelines. Use Amazon SageMaker Studio Classic to create the model deployment pipelines.
Answer: (SHOW ANSWER)

NEW QUESTION: 96
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: (SHOW ANSWER)
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.

NEW QUESTION: 97
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: (SHOW ANSWER)
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.

NEW QUESTION: 98
An ML engineer is using Amazon SageMaker Canvas to build a custom ML model from an imported dataset.
The model must make continuous numeric predictions based on 10 years of data.
Which metric should the ML engineer use to evaluate the model's performance?
A. Accuracy
B. InferenceLatency
C. Area Under the ROC Curve (AUC)
D. Root Mean Square Error (RMSE)
Answer: D (LEAVE A REPLY)
This is a regression problem, where the target variable is continuous and numeric. AWS documentation clearly states that classification metrics such as accuracy and AUC are not appropriate for regression models.
Root Mean Square Error (RMSE) measures the square root of the average squared differences between predicted and actual values. RMSE penalizes larger errors more heavily, making it especially useful when large prediction errors are costly or undesirable.
SageMaker Canvas automatically selects regression metrics such as RMSE and MAE when building regression models. RMSE is widely used for time-based and numeric prediction problems, especially when evaluating long historical datasets.
Inference latency measures system performance, not model accuracy.
Therefore, Option D is the correct and AWS-verified answer.

NEW QUESTION: 99
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 original baseline data had a data quality issue of missing values.
C. The model was not sufficiently complex to capture all the patterns in the original baseline data.
D. Incorrect ground truth labels were provided to Model Monitor during the calculation of the baseline.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 100
A company is developing an ML model by using Amazon SageMaker AI. The company must monitor bias in the model and display the results on a dashboard. An ML engineer creates a bias monitoring job.
How should the ML engineer capture bias metrics to display on the dashboard?
A. Capture AWS CloudTrail metrics from SageMaker Clarify.
B. Capture Amazon CloudWatch metrics from SageMaker Clarify.
C. Capture SageMaker Model Monitor metrics from Amazon EventBridge.
D. Capture SageMaker Model Monitor metrics from Amazon SNS.
Answer: B (LEAVE A REPLY)
Amazon SageMaker Clarify is the AWS service used to detect and quantify bias and fairness metrics in ML models. When bias monitoring jobs run, Clarify publishes bias metrics directly to Amazon CloudWatch.
CloudWatch metrics can be visualized using CloudWatch dashboards or integrated into other monitoring tools, making them ideal for real-time or periodic bias reporting.
CloudTrail logs API activity and does not capture ML metrics. EventBridge and SNS are used for event routing and notifications, not metric visualization.
AWS documentation explicitly states that Clarify bias metrics are emitted to Amazon CloudWatch, which is the correct source for dashboards.
Therefore, Option B is the correct and AWS-verified answer.

NEW QUESTION: 101
A company needs to use Retrieval Augmented Generation (RAG) to supplement an open source large language model (LLM) that runs on Amazon Bedrock. The company's data for RAG is a set of documents in an Amazon S3 bucket. The documents consist of .csv files and .docx files.
Which solution will meet these requirements with the LEAST operational overhead?
A. Fine-tune an existing LLM by using an AutoML job in Amazon SageMaker. Configure the S3 bucket as a data source for the AutoML job. Deploy the LLM to a SageMaker endpoint. Use the endpoint to perform RAG queries.
B. Create a knowledge base for Amazon Bedrock. Configure a data source that references the S3 bucket. Use the Amazon Bedrock API to perform RAG queries.
C. Convert the data into vectors. Store the data in an Amazon Neptune database. Connect the database to Amazon Bedrock. Call the Amazon Bedrock API to perform RAG queries.
D. Create a pipeline in Amazon SageMaker Pipelines to generate a new model. Call the new model from Amazon Bedrock to perform RAG queries.
Answer: (SHOW ANSWER)

NEW QUESTION: 102
A data scientist uses logistic regression to build a fraud detection model. While the model accuracy is 99%, 90% of the fraud cases are not detected by the model.
What action will definitively help the model detect more than 10% of fraud cases?
A. Using undersampling to balance the dataset
B. Decreasing the class probability threshold
C. Using regularization to reduce overfitting
D. Using oversampling to balance the dataset
Answer: B (LEAVE A REPLY)

Decreasing the class probability threshold makes the model more sensitive and, therefore, marks more cases as the positive class, which is fraud in this case. This will increase the likelihood of fraud detection. However, it comes at the price of lowering precision.

NEW QUESTION: 103
A company must install a custom script on any newly created Amazon SageMaker AI notebook instances.
Which solution will meet this requirement with the LEAST operational overhead?
A. Create a lifecycle configuration script to install the custom script when a new SageMaker AI notebook is created. Attach the lifecycle configuration to every new SageMaker AI notebook as part of the creation steps.
B. Create a custom Amazon Elastic Container Registry (Amazon ECR) image that contains the custom script. Push the ECR image to a Docker registry. Attach the Docker image to a SageMaker Studio domain. Select the kernel to run as part of the SageMaker AI notebook.
C. Create a custom package index repository. Use AWS CodeArtifact to manage the installation of the custom script. Set up AWS PrivateLink endpoints to connect CodeArtifact to the SageMaker AI instance. Install the script.
D. Store the custom script in Amazon S3. Create an AWS Lambda function to install the custom script on new SageMaker AI notebooks. Configure Amazon EventBridge to invoke the Lambda function when a new SageMaker AI notebook is initialized.
Answer: (SHOW ANSWER)
AWS recommends lifecycle configuration scripts as the simplest and most direct way to customize Amazon SageMaker Notebook Instances at creation time. Lifecycle configurations run automatically when a notebook instance is created or started, allowing scripts, packages, and system dependencies to be installed without manual intervention.
This approach is fully supported, requires no additional infrastructure, and integrates directly with the notebook creation workflow. The script can be reused across notebooks, ensuring consistency.
Options B, C, and D introduce unnecessary complexity, such as container management, private package repositories, or event-driven orchestration.
Therefore, lifecycle configuration scripts provide the least operational overhead solution.

NEW QUESTION: 104
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. Deploy models by using AWS Lambda functions.
C. Use PyTorch or TensorFlow with the distributed training option.
D. Use Amazon SageMaker Ground Truth for data labeling.
E. Use AWS Trainium instances for training.
Answer: A,E (LEAVE A REPLY)

NEW QUESTION: 105
A company uses Amazon SageMaker for its ML process. A compliance audit discovers that an Amazon S3 bucket for training data uses server-side encryption with S3 managed keys (SSE- S3).
The company requires customer managed keys. An ML engineer changes the S3 bucket to use server-side encryption with AWS KMS keys (SSE-KMS). The ML engineer makes no other configuration changes.
After the change to the encryption settings, SageMaker training jobs start to fail with AccessDenied errors.
What should the ML engineer do to resolve this problem?
A. Update the IAM policy that is attached to the user that created the training jobs. Include the kms:CreateGrant permission.
B. Update the IAM policy that is attached to the execution role for the training jobs. Include the s3:ListBucket and s3:GetObject permissions.
C. Update the IAM policy that is attached to the execution role for the training jobs. Include the kms:Encrypt and kms:Decrypt permissions.
D. Update the S3 bucket policy that is attached to the S3 bucket. Set the value of the aws:SecureTransport condition key to True.

Answer: C (LEAVE A REPLY)

NEW QUESTION: 106
A company is uploading thousands of PDF policy documents into Amazon S3 and Amazon Bedrock Knowledge Bases. Each document contains structured sections. Users often search for a small section but need the full section context. The company wants accurate section-level search with automatic context retrieval and minimal custom coding.
Which chunking strategy meets these requirements?
A. Hierarchical
B. Maximum tokens
C. Semantic
D. Fixed-size
Answer: A (LEAVE A REPLY)
AWS Bedrock Knowledge Bases support multiple chunking strategies to optimize retrieval quality.
Hierarchical chunking is specifically designed for structured documents such as PDFs with headings, sections, and subsections.
Hierarchical chunking allows fine-grained retrieval at the subsection level while automatically preserving parent section context. This ensures that when a small portion is retrieved, the surrounding section is also provided to the foundation model for better understanding.
Fixed-size and maximum-token chunking can split content arbitrarily, breaking semantic and structural boundaries. Semantic chunking focuses on meaning but does not guarantee structured context preservation without additional logic.
AWS documentation highlights hierarchical chunking as the preferred strategy when documents are structured and contextual integrity is required.
Therefore, Option A is the correct and AWS-aligned solution.

NEW QUESTION: 107
A company runs an Amazon SageMaker AI domain in a public subnet of a newly created VPC. The network is configured properly, and ML engineers can access the SageMaker AI 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 ACL 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 (LEAVE A REPLY)
In AWS networking, security groups are stateful and allow-only, meaning they cannot explicitly deny traffic.
As a result, Option A is invalid. Network ACLs (NACLs), on the other hand, are stateless and support both allow and deny rules, making them the correct mechanism for blocking traffic from specific IP addresses.
Because the SageMaker AI domain is deployed in a public subnet, inbound traffic reaches the subnet before it reaches the resource. AWS documentation states that NACLs are evaluated at the subnet level and are ideal for implementing IP-based blocking rules.
Route tables control routing paths, not traffic filtering, so Option D is incorrect. Option C is unrelated to network security and does not block traffic.

AWS best practices clearly recommend using network ACL deny rules when an explicit block is required for a specific IP address at the subnet boundary.
Therefore, Option B is the correct and AWS-aligned solution.

NEW QUESTION: 108
A company is building a conversational AI assistant on Amazon Bedrock. The company is using Retrieval Augmented Generation (RAG) to reference the company's internal knowledge base. The AI assistant uses the Anthropic Claude 4 foundation model (FM).
The company needs a solution that uses a vector embedding model, a vector store, and a vector search algorithm.
Which solution will develop the AI assistant with the LEAST development effort?
A. Use Amazon Kendra Experience Builder.
B. Use Amazon Aurora PostgreSQL with the pgvector extension.
C. Use Amazon RDS for PostgreSQL with the pgvector extension.
D. Use the AWS Glue Data Catalog metadata repository.
Answer: (SHOW ANSWER)
Amazon Kendra Experience Builder provides a fully managed, low-code solution for building conversational search and question-answering applications. AWS documentation states that Kendra natively supports semantic search, vector embeddings, and vector-based retrieval, making it well suited for RAG-style applications with minimal development effort.
When integrated with Amazon Bedrock, Kendra can act as the retrieval layer, handling document ingestion, indexing, embedding generation, and relevance ranking automatically. This eliminates the need to manually manage embedding models, vector databases, and search logic.
Options B and C require custom schema design, vector indexing, query logic, and operational management of PostgreSQL instances. Although pgvector supports vector search, it significantly increases development and maintenance effort. Option D is unrelated to vector search and is used only for metadata cataloging.
AWS explicitly positions Amazon Kendra as the fastest way to build enterprise-grade conversational assistants that integrate with foundation models.
Therefore, Option A is the correct and most AWS-aligned solution.

NEW QUESTION: 109
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 (LEAVE A REPLY)
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.

NEW QUESTION: 110
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 each user's IAM profile. Configure AWS Budgets to send an alert when the threshold is reached.
C. 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.
D. Add resource tagging by editing each user's IAM profile. Configure AWS Cost Explorer to send an alert when the threshold is reached.
Answer: C (LEAVE A REPLY)

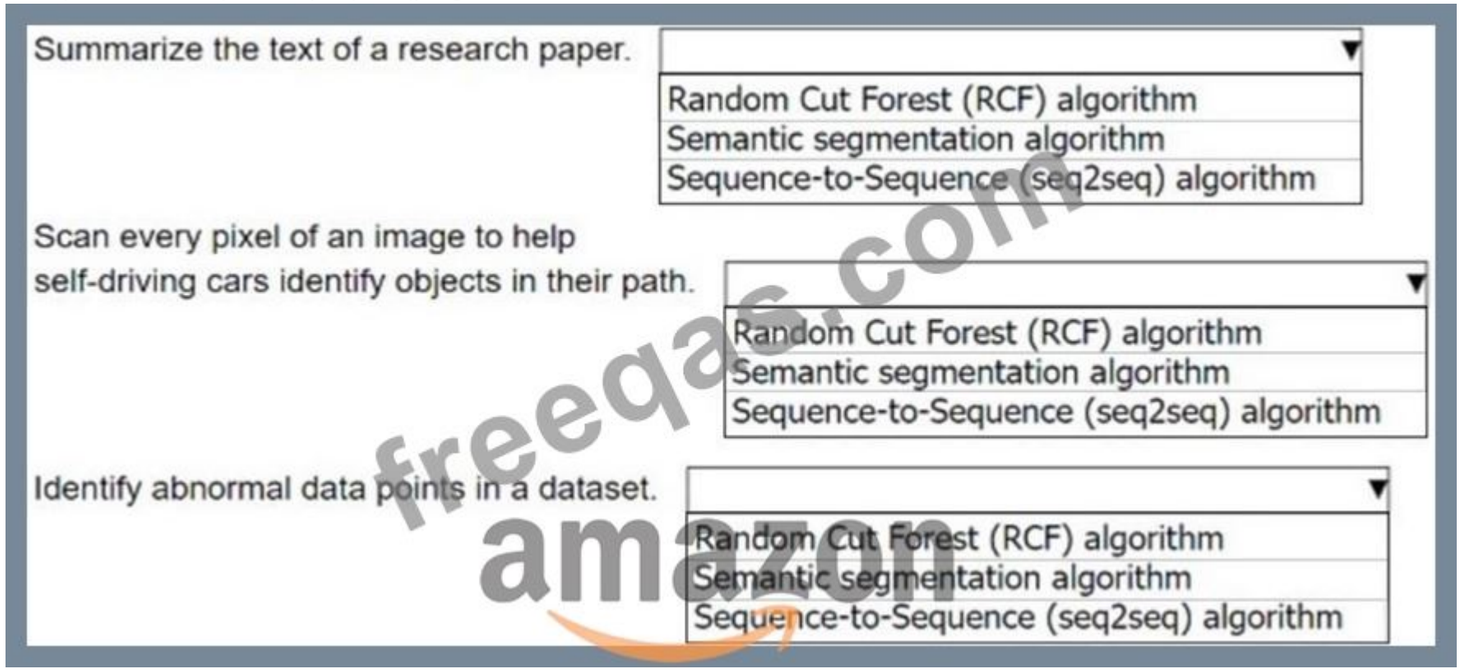
NEW QUESTION: 111
A company has a Retrieval Augmented Generation (RAG) application that uses a vector database to store embeddings of documents. The company must migrate the application to AWS and must implement a solution that provides semantic search of text files. The company has already migrated the text repository to an Amazon S3 bucket.
Which solution will meet these requirements?
A. Use the Amazon Kendra S3 connector to ingest the documents from the S3 bucket into Amazon Kendra. Query Amazon Kendra 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 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.
D. Use an Amazon Textract asynchronous job to ingest the documents from the S3 bucket. Query Amazon Textract to perform the semantic searches.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 112
A company shares Amazon SageMaker Studio notebooks that are accessible through a VPN.
The company must enforce access controls to prevent malicious actors from exploiting presigned URLs to access the notebooks.
Which solution will meet these requirements?
A. Set up Studio client IP validation by using the aws:sourceIp IAM policy condition.
B. Set up Studio client role endpoint validation by using the aws:PrimaryTag IAM policy condition.
C. Set up Studio client user endpoint validation by using the aws:PrincipalTag IAM policy condition.
D. Set up Studio client VPC validation by using the aws:sourceVpc IAM policy condition.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 113
A digital media entertainment company needs real-time video content moderation to ensure compliance during live streaming events.
Which solution will meet these requirements with the LEAST operational overhead?
A. Use Amazon Rekognition and AWS Lambda to extract and analyze the metadata from the videos' image frames.
B. Use Amazon Rekognition and a large language model (LLM) hosted on Amazon Bedrock to extract and analyze the metadata from the videos' image frames.
C. Use Amazon SageMaker AI to extract and analyze the metadata from the videos' image frames.
D. Use Amazon Transcribe and Amazon Comprehend to extract and analyze the metadata from the videos' image frames.
Answer: (SHOW ANSWER)
For real-time video content moderation with minimal operational overhead, AWS documentation recommends using fully managed, purpose-built AI services. Amazon Rekognition provides real-time video analysis capabilities, including content moderation, unsafe content detection, and label recognition for live video streams.
By integrating Rekognition with AWS Lambda, the company can automatically process video frames, extract moderation metadata, and take immediate action (such as flagging or stopping a stream) without managing servers, models, or infrastructure. This serverless architecture scales automatically and minimizes operational complexity.
Option B introduces unnecessary complexity. While Amazon Bedrock LLMs are powerful, they are not required for image-based moderation tasks that Rekognition already handles natively.
Option C is incorrect because using Amazon SageMaker would require model training, endpoint management, and scaling, significantly increasing operational overhead.
Option D is incorrect because Amazon Transcribe and Amazon Comprehend are designed for audio and text analysis, not image or video frame moderation.

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

NEW QUESTION: 114
Hotspot Question
An ML engineer must choose the appropriate Amazon SageMaker algorithm to solve specific AI problems.
Select the correct SageMaker built-in algorithm from the following list for each use case. Each algorithm should be selected one time.
- Random Cut Forest (RCF) algorithm
- Semantic segmentation algorithm
- Sequence-to-Sequence (seq2seq) algorithm



Answer:

Summarize the text of a research paper.

- Random Cut Forest (RCF) algorithm
- Semantic segmentation algorithm
- Sequence-to-Sequence (seq2seq) algorithm

Scan every pixel of an image to help self-driving cars identify objects in their path.

- Random Cut Forest (RCF) algorithm
- Semantic segmentation algorithm
- Sequence-to-Sequence (seq2seq) algorithm

Identify abnormal data points in a dataset.

- Random Cut Forest (RCF) algorithm
- Semantic segmentation algorithm
- Sequence-to-Sequence (seq2seq) algorithm

NEW QUESTION: 115
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 (LEAVE A REPLY)
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.

NEW QUESTION: 116
A company is using Amazon SageMaker AI to develop a credit risk assessment model. During model validation, the company finds that the model achieves 82% accuracy on the validation data. However, the model achieved 99% accuracy on the training data. The company needs to address the model accuracy issue before deployment.
Which solution will meet this requirement?
A. Add more dense layers to increase model complexity. Implement batch normalization. Use early stopping during training.

B. Implement dropout layers. Use L1 or L2 regularization. Perform k-fold cross-validation.
C. Use principal component analysis (PCA) to reduce the feature dimensionality. Decrease model layers.
Implement cross-entropy loss functions.
D. Augment the training dataset. Remove duplicate records from the training dataset. Implement stratified sampling.
Answer: B (LEAVE A REPLY)
The large gap between training accuracy (99%) and validation accuracy (82%) is a textbook case of overfitting. The model has learned patterns that fit the training data extremely well but do not generalize to unseen data.
AWS ML best practices recommend regularization techniques to address overfitting. Dropout layers randomly deactivate neurons during training, preventing the network from relying too heavily on specific paths. L1 and L2 regularization penalize large weights, reducing model complexity and improving generalization. k-fold cross-validation provides a more robust evaluation by training and validating the model across multiple data splits.
Option A increases complexity, which would worsen overfitting. Option C mixes valid ideas (dimensionality reduction) with unrelated changes (loss function choice) and is less targeted. Option D focuses on data quality but does not directly address model variance.
Therefore, implementing dropout, regularization, and k-fold cross-validation is the correct solution.

NEW QUESTION: 117
An ML engineer needs to train a supervised deep learning model. The available dataset is a large number of unlabeled images that only employees should access. The ML engineer needs to implement a solution that labels the dataset with the highest possible accuracy. Which combination of steps should the ML engineer take to meet these requirements? (Choose two.)
A. Use Amazon Rekognition to automatically label the dataset.
B. Train the deep learning model directly on the raw data. Let the model infer the labels by itself.
C. Use Amazon SageMaker Ground Truth to create an annotation job that specifies the labeling task and requirements.
D. Set up workforce teams to access a private workforce to run and review the annotation job created by Amazon SageMaker Ground Truth.
E. Use Amazon Mechanical Turk to complete the annotation job created by Amazon SageMaker Ground Truth.
Answer: C,D (LEAVE A REPLY)
To achieve the highest labeling accuracy with controlled employee-only access, the ML engineer should use Amazon SageMaker Ground Truth to define the annotation job and then assign it to a private workforce of employees for labeling and review. This ensures high-quality, secure labeling restricted to authorized personnel.

NEW QUESTION: 118
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 (LEAVE A REPLY)
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.

NEW QUESTION: 119
A company has several teams that have developed separate prediction models on their own laptops. The teams developed the models by using Python with scikit-learn and TensorFlow frameworks. The company must rebuild the models and must integrate the models into an ML infrastructure that the company manages by using Amazon SageMaker. The company also must incorporate the

models into a model registry.
Which solution will meet these requirements with the LEAST operational overhead?
A. Export the models from the laptops to an Amazon S3 bucket. Use an Amazon API Gateway REST API and AWS Lambda functions with SageMaker endpoints to access the models. Register the models in the SageMaker Model Registry.
B. Import the models into the SageMaker Model Registry. Use SageMaker to run the imported models.
C. Use code from the laptops to create containers for the models. Use the bring your own container (BYOC) functionality of SageMaker to import and use the models. Register the models in the SageMaker Model Registry.
D. Import the Python-based models into SageMaker. Rebuild the scikit-learn and TensorFlow models in SageMaker. Register all the models in the SageMaker Model Registry.
Answer: D (LEAVE A REPLY)
The least operational overhead comes from directly importing the scikit-learn and TensorFlow models into SageMaker, rebuilding them using the respective prebuilt SageMaker frameworks, and then registering them in the SageMaker Model Registry. This leverages managed framework containers provided by SageMaker, avoids custom container management, and integrates seamlessly with the registry.

NEW QUESTION: 120
A machine learning engineer is preparing a data frame for a supervised learning task with the Amazon SageMaker Linear Learner algorithm. The ML engineer notices the target label classes are highly imbalanced and multiple feature columns contain missing values. The proportion of missing values across the entire data frame is less than 5%.
What should the ML engineer do to minimize bias due to missing values?
A. Replace each missing value by the mean or median across non-missing values in same row.
B. Delete observations that contain missing values because these represent less than 5% of the data.
C. Replace each missing value by the mean or median across non-missing values in the same column.
D. For each feature, approximate the missing values using supervised learning based on other features.
Answer: D (LEAVE A REPLY)
Use supervised learning to predict missing values based on the values of other features. Different supervised learning approaches might have different performances, but any properly implemented supervised learning approach should provide the same or better approximation than mean or median approximation, as proposed in responses A and C. Supervised learning applied to the imputation of missing values is an active field of research.