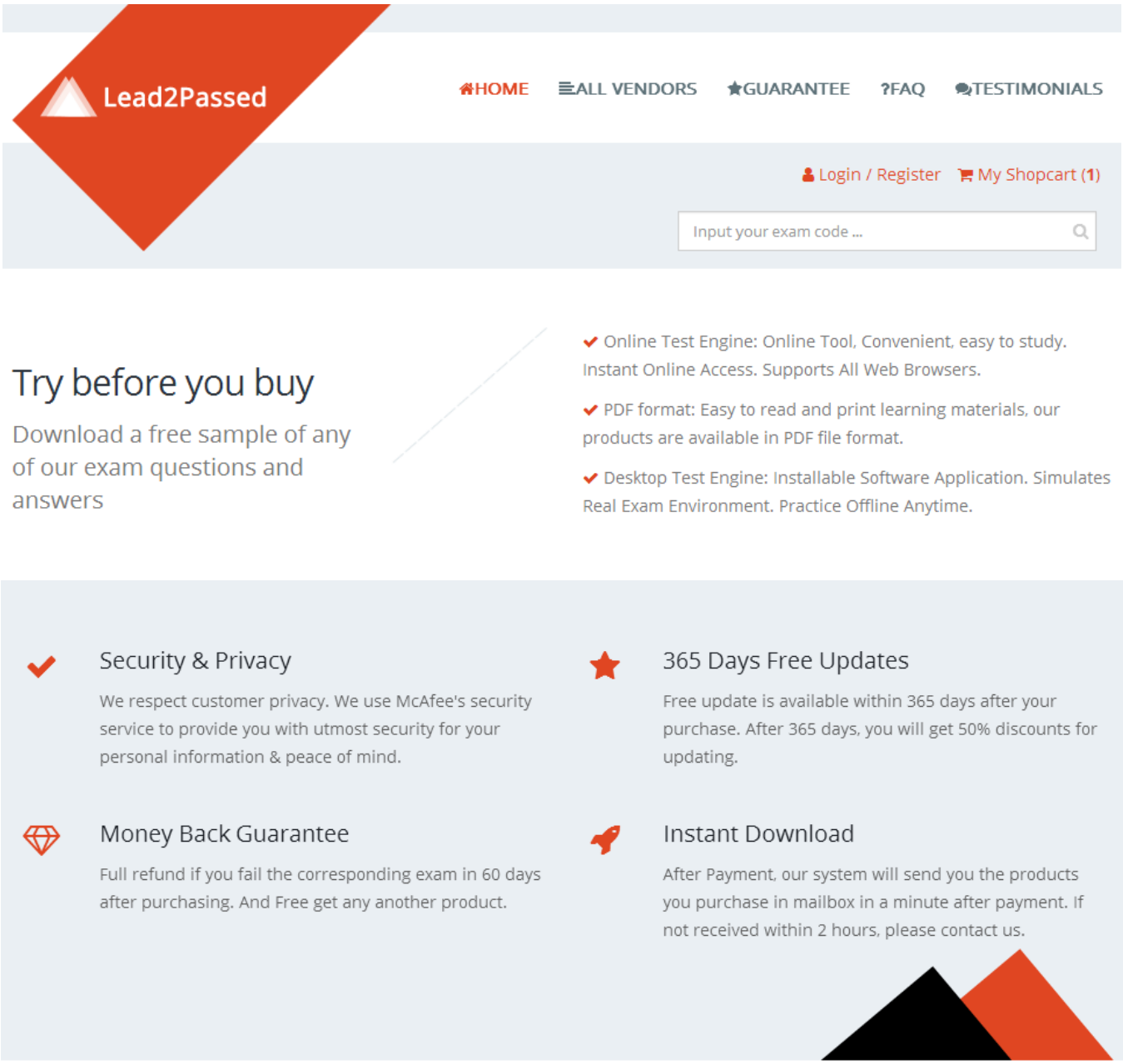
Lead2Passed



http://www.lead2passed.com

Valid Certification Exam Dumps Materials and Study Guide - Lead2Passed

Exam : **MLA-C01-JPN**

Title : AWS Certified Machine Learning Engineer - Associate (MLA-C01日本語版)

Vendor : Amazon

Version : DEMO

QUESTION NO: 1
MLエンジニアがAmazon SageMaker
AIでトレーニングジョブを実行したいと考えています。このトレーニングジョブでは、複数のGPUを使用してニューラルネットワークをトレーニングします。トレーニングデータセットはParquet形式で保存されています。
ML エンジニアは、Parquet データセットに、SageMaker AI トレーニング
インスタンスのメモリに収まらないほど大きなファイルが含まれていることを発見しました。
どの解決策がメモリの問題を解決するでしょうか?
A. Amazon Elastic Block Store (Amazon EBS) プロビジョンド IOPS SSD
ボリュームをインスタンスに接続します。
ファイルを EBS ボリュームに保存します。
B. Amazon EMR 上の Apache Spark を使用して Parquet
ファイルを再パーティション化します。再パーティション化されたファイルをトレーニングジョブに使用します。
C. インスタンスタイプを、トレーニング
ジョブに十分なメモリを備えたメモリ最適化インスタンスに変更します。
D. 複数のインスタンスで SageMaker AI 分散データ並列処理 (SMDDP)
ライブラリを使用して、メモリ使用量を分割します。
Answer: B
Explanation:
The issue is caused by oversized Parquet files that cannot be efficiently read into memory during training. The most effective and scalable solution is to repartition the dataset into smaller Parquet files.
AWS best practices for large-scale ML training recommend optimizing data layout, not simply increasing memory. By using Apache Spark on Amazon EMR, the ML engineer can repartition the Parquet files into smaller chunks that can be streamed and processed efficiently by SageMaker training jobs.
Attaching EBS volumes (Option A) increases storage capacity but does not solve in-memory constraints.
Changing to memory-optimized instances (Option C) increases cost and does not address long-term scalability. SMDDP (Option D) distributes gradients and computation, not dataset file sizes.
Therefore, repartitioning the Parquet files is the correct solution.

QUESTION NO: 2
機械学習エンジニアは、分析のためにAmazon
S3からデータを利用、準備、ロードしたいと考えています。そのため、データのスキーマを検出し、メタデータを保存するために、抽出、変換、ロード(ETL)ジョブを実行する必要があります。
これらの要件を最も少ない手作業で満たすソリューションはどれでしょうか？
A. AWS Glue を使用して ETL
ジョブを実行します。このジョブを使用してスキーマを検出し、関連するメタデータを
AWS Glue データカタログに保存します。
B. ETL ジョブを実行する Amazon SageMaker Data Wrangler

フローを作成します。このジョブを使用してスキーマを検出し、関連するメタデータを S3 バケットに保存します。

C. Amazon AthenaとAWS Step Functionsを統合してETLパイプラインを作成します。このパイプラインを使用してETLジョブを実行し、スキーマを検出して関連するメタデータをS3バケットに保存します。

D. scikit-learnライブラリを含むAmazon EC2インスタンスを起動し、ETLジョブを実行します。このジョブを使用してスキーマを検出し、関連するメタデータをAmazon Redshiftに保存します。

Answer: A

Explanation:

Option A is correct because AWS Glue is the AWS-native managed ETL service built specifically to discover schema, run ETL jobs, and store metadata in the AWS Glue Data Catalog. AWS documentation states that Glue crawlers can automatically discover and catalog new or updated data sources, and that the Data Catalog automatically captures and manages schema metadata. This directly matches the requirement to run an ETL job on data in Amazon S3, discover the schema, and store the metadata with the least manual effort.

AWS Glue is also the lowest-effort answer because the service is managed and purpose-built for this workflow. The Glue Data Catalog serves as a persistent metadata repository, and AWS documents that crawlers infer schema information and integrate it into the catalog automatically. That means the ML engineer does not need to build custom schema inference logic or manually maintain metadata storage. This is exactly the kind of manual work the question is trying to avoid.

The other options are not as good. SageMaker Data Wrangler is primarily for visual data preparation and feature engineering, not for running a managed ETL-plus-catalog workflow with schema stored in a metadata catalog. Athena with Step Functions would require assembling more custom orchestration and still does not naturally replace the Glue Data Catalog workflow. Launching an EC2 instance introduces the highest operational overhead and does not align with the requirement for least manual effort. Therefore, the best verified AWS-docs answer is A, because AWS Glue combines ETL, schema discovery, and metadata cataloging in one managed service.

QUESTION NO: 3

ある企業では、文書の埋め込み情報をベクターデータベースに保存するRetrieval Augmented Generation（RAG）アプリケーションを使用しています。このアプリケーションをAWSに移行し、テキストファイルのセマンティック検索を提供するソリューションを実装する必要があります。テキストリポジトリは既にAmazon S3バケットに移行済みです。

これらの要件を満たすソリューションはどれでしょうか?

A. AWS Batch ジョブを使用してファイルを処理して埋め込みを生成します。AWS Glue を使用して埋め込みを保存します。SQL クエリを使用してセマンティック検索を実行します。

B. Amazon SageMakerのカスタムノートブックを使用して、カスタムスクリプトを実行し、埋め込みを生成します。埋め込みはSageMaker Feature Storeに保存します。SQLクエリを使用してセマンティック検索を実行します。

C. Amazon Kendra S3 コネクタを使用して、S3 バケットから Amazon Kendra にドキュメントを取り込み、Amazon Kendra にクエリを実行してセマンティック検索を実行します。

D. Amazon Textract の非同期ジョブを使用して、S3 バケットからドキュメントを取り込み、Amazon Textract にクエリを実行してセマンティック検索を実行します。

Answer: C

Explanation:

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

QUESTION NO: 4

ある企業がAmazon SageMakerで機械学習モデルをトレーニングしました。本番環境で推論を提供するために、このモデルをホストする必要があります。

モデルは高可用性を備え、最小限のレイテンシで応答する必要があります。各リクエストのサイズは1KB～3MBです。モデルは日中に予測不可能なリクエストの集中的な発生を経験することになります。推論は需要の変化に比例して適応する必要があります。

これらの要件を満たすために、企業はどのようにモデルを本番環境に導入すべきでしょうか？

A. SageMaker リアルタイム推論エンドポイントを作成します。自動スケーリングを設定します。既存のモデルを表示するようにエンドポイントを設定します。

B. Amazon Elastic Container Service (Amazon ECS) クラスターにモデルをデプロイします。ECS クラスターの CPU に基づいてスケジュール設定された ECS スケーリングを使用します。

C. Amazon Elastic Kubernetes Service (Amazon EKS) クラスターに SageMaker Operator をインストールします。Amazon EKS にモデルをデプロイします。メモリメトリクスに基づいてレプリカをスケーリングするように、水平ポッドオートスケーリングを設定します。

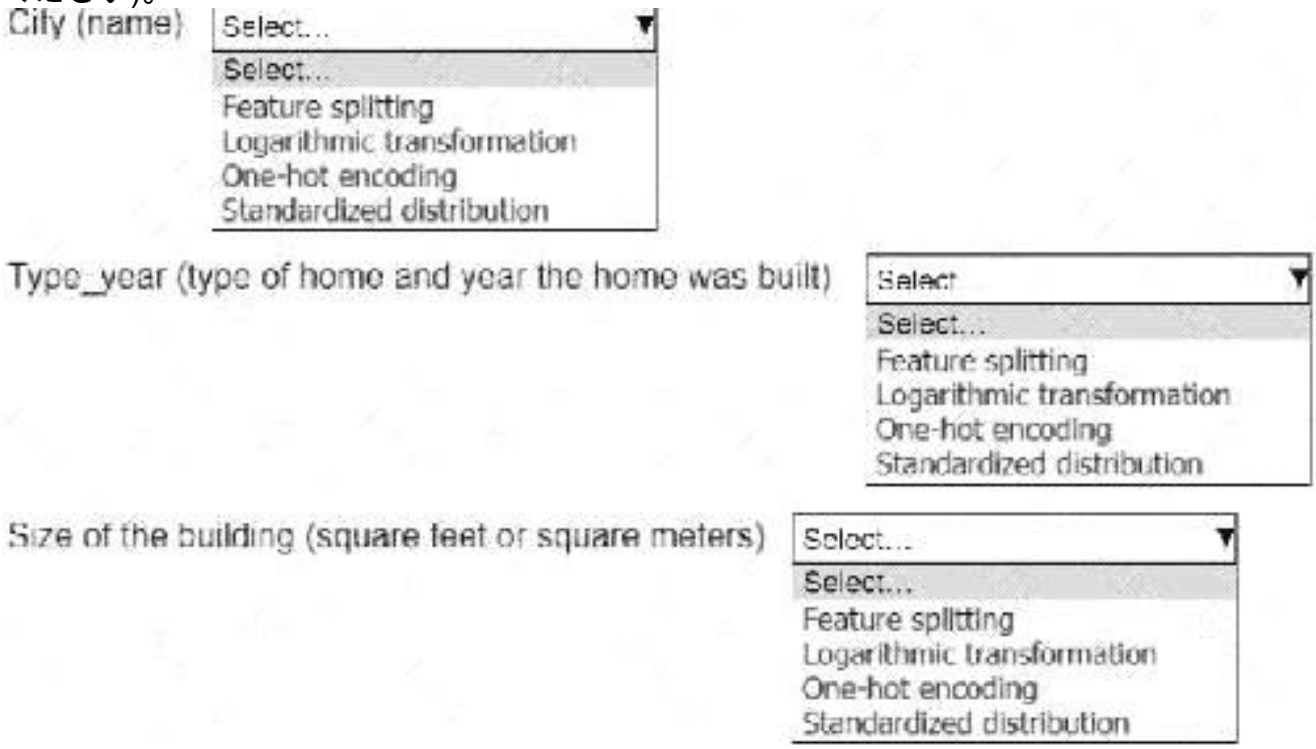
D. 推論には、Application Load Balancer (ALB) の背後にあるスポットフリートでスポットインスタンスを使用します。自動スケーリングのメトリクスとして、ALBRequestCountPerTarget メトリクスを使用します。

Answer: A

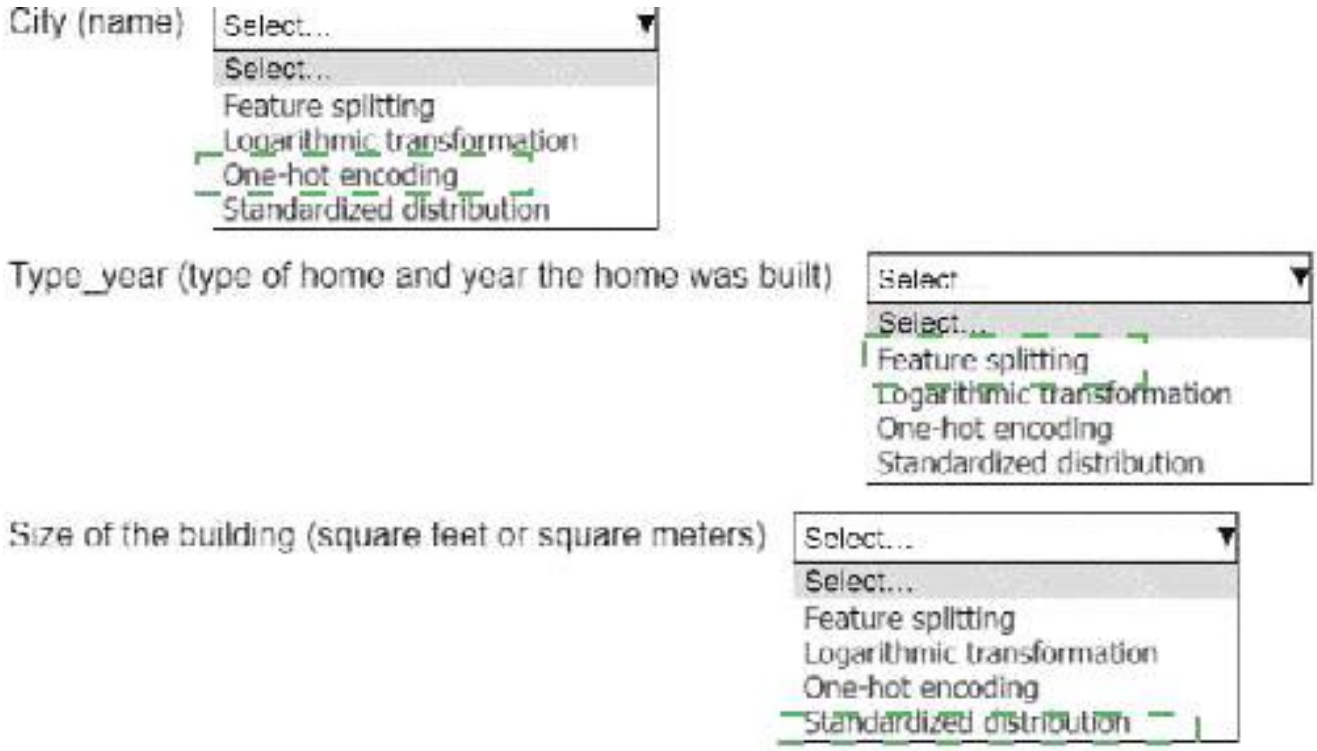
Explanation:

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

QUESTION NO: 5
MLエンジニアは、同規模の住宅の価格を予測するMLモデルの開発に取り組んでいます。このモデルは、複数の特徴量に基づいて予測を行います。MLエンジニアは、以下の特徴量エンジニアリング手法を用いて住宅の価格を推定します。
* 機能分割
* 対数変換
* ワンホットエンコーディング
* 標準化された配布
以下の特徴量リストについて、正しい特徴量エンジニアリング手法を選択してください。各特徴量エンジニアリング手法は、1回選択するか、全く選択しないでください(3つ選択してください)。



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.

QUESTION NO: 6
ある企業は、機械学習モデルのベンダーから定期的に新しいトレーニングデータを受け取ります。ベンダーは、3～4日ごとに、クリーンアップおよび準備されたデータを同社のAmazon S3バケットに配信します。
同社には、モデルを再トレーニングするためのAmazon SageMakerパイプラインがあります。MLエンジニアは、S3バケットに新しいデータがアップロードされたときにパイプラインを実行するソリューションを実装する必要があります。最も少ない運用労力でこれらの要件を満たすソリューションはどれでしょうか?
A. S3 ライフサイクル ルールを作成して、データを SageMaker トレーニング

インスタンスに転送し、トレーニングを開始します。
B. S3バケットをスキャンするAWS
Lambda関数を作成します。新しいデータがアップロードされたときにパイプラインを開始するようにLambda関数をプログラムします。
C. S3アップロードに一致するイベントパターンを持つAmazon
EventBridgeルールを作成します。ルールのターゲットとしてパイプラインを設定します。
D.
新しいデータがアップロードされたときにパイプラインをオーケストレーションするには、Amazon Managed Workflows for Apache Airflow (Amazon MWAA) を使用します。
Answer: C
Explanation:
Using Amazon EventBridge with 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.

QUESTION NO: 7
ある企業では、Amazon EMR
クラスターを使用して機械学習モデルのデータ取り込みプロセスを実行しています。機械学習エンジニアは、処理時間が増加していることに気付きました。
最もコスト効率よく処理時間を短縮できるソリューションはどれですか?
A. スポットインスタンスを使用してプライマリノードの数を増やします。
B. スポットインスタンスを使用してコアノードの数を増やします。
C. スポットインスタンスを使用してタスクノードの数を増やします。
D. オンデマンドインスタンスを使用して、コアノードの数を増やします。
Answer: C
Explanation:
Amazon EMR clusters consist of primary, core, and task nodes, each with a distinct role. The primary node manages the cluster, core nodes store data and run tasks, and task nodes only run tasks without storing data.
AWS documentation recommends using task nodes for scaling compute capacity when workloads are compute-intensive, such as data ingestion and transformation pipelines.
To reduce processing time cost-effectively, AWS strongly advises using Spot Instances for task nodes. Spot Instances provide the same compute capacity as On-Demand Instances but at a significantly reduced cost, often up to 90% lower. Because task nodes do not store HDFS data, they can be safely interrupted without risking data loss.
Increasing the number of primary nodes is not supported by EMR and would not improve performance.
Increasing core nodes affects both storage and compute and is more expensive, especially when using On- Demand Instances. Option D is therefore the least cost-effective.
AWS EMR best practices explicitly state that scaling out with Spot task nodes is the preferred way to improve performance for transient, parallel workloads such as ETL, ingestion, and feature preparation.
Therefore, Option C is the most cost-effective and AWS-recommended solution.

QUESTION NO: 8
機械学習エンジニアは、Amazon SageMaker
上で機械学習モデルをトレーニングし、防爆型テレビ映像から自動車事故を検出しました。
また、SageMaker Data Wrangler
を使用して、事故画像と非事故画像のトレーニングデータセットを作成しました。
モデルはトレーニングと検証では良好なパフォーマンスを示しました。しかし、様々なカメラからの画像品質のばらつきにより、本番環境ではパフォーマンスが低下しています。
どのソリューションが最も短い時間でモデルの精度を向上させるでしょうか?
A. すべてのカメラからさらに画像を収集します。Data Wrangler
を使用して新しいトレーニングデータセットを準備します。
B. Data
Wranglerの破損画像変換を使用して、トレーニングデータセットを再作成します。インパルスノイズオプションを指定します。
C. Data Wrangler
の画像コントラスト強化変換を使用して、トレーニングデータセットを再作成します。ガンマコントラストオプションを指定します。
D. Data Wrangler
の画像サイズ変更変換を使用して、トレーニングデータセットを再作成します。すべての画像を同じサイズにトリミングします。
Answer: B
Explanation:
The model is underperforming in production due to variations in image quality from different cameras. Using the corrupt image transform with the impulse noise option in SageMaker Data Wrangler simulates real-world noise and variations in the training dataset. This approach helps the model become more robust to inconsistencies in image quality, improving its accuracy in production without the need to collect and process new data, thereby saving time.

QUESTION NO: 9
機械学習エンジニアが、2つのクラスのうち1つでパフォーマンスが低い画像分類モデルを調整しています。パフォーマンスが低いクラスは、トレーニングデータセットのごく一部を占めています。
どのソリューションがモデルのパフォーマンスを向上させるでしょうか?
A. 精度を最適化します。あまり一般的ではない画像には画像拡張を使用します。
B. F1スコアを最適化します。あまり一般的でない画像には画像拡張を使用します。
C. 精度を最適化します。SMOTEを使用して合成画像を生成します。
D. F1スコアを最適化します。SMOTEを用いて合成画像を生成します。
Answer: B
Explanation:
This scenario describes a severely imbalanced classification problem. In such cases, accuracy is a misleading metric, because the model can achieve high accuracy by predicting only the majority class.
AWS ML best practices recommend using F1 score (or precision/recall) when evaluating imbalanced datasets.

The F1 score balances false positives and false negatives, making it ideal for assessing minority-class performance.
For image data, image augmentation (rotations, flips, crops, color jitter) is the preferred technique to increase minority-class representation. SMOTE is designed for tabular data and is not suitable for image pixel data.
Therefore, the correct solution is to optimize for F1 score and apply image augmentation.
Thus, Option B is the correct and AWS-aligned answer.

QUESTION NO: 10
MLエンジニアが住宅とアパートの価格を予測するモデルを構築しています。このモデルは3つの特徴量を使用しています。
平方メートル、価格、築年数。データセットには10,000行のデータが含まれています。データには、大きなマンション1棟と非常に小さなアパート1棟のデータポイントが含まれています。
ML
エンジニアは、モデルが一般的な住宅やアパートに対して正確な予測を生成することを確認するために、データセットの前処理を実行する必要があります。
これらの要件を満たすソリューションはどれでしょうか?
A. 外れ値を削除し、Square Meters 変数に対して対数変換を実行します。
B. 外れ値を保持し、Square Meters 変数に対して正規化を実行します。
C. 外れ値を削除し、Square Meters
変数に対してワンホットエンコーディングを実行します。
D. 外れ値を保持し、Square Meters
変数に対してワンホットエンコーディングを実行します。
Answer: A
Explanation:
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.

QUESTION NO: 11
ある金融会社は、外部プロバイダーから大量のリアルタイム市場データストリームを受信しています。ストリームは毎秒数千件のJSONレコードで構成されています。
同社は、異常なデータポイントを識別するために、AWS
上にスケーラブルなソリューションを実装する必要があります。
最も少ない運用オーバーヘッドでこれらの要件を満たすソリューションはどれでしょうか?
A. Amazon Kinesis Data Streams にリアルタイムデータを取り込みます。Amazon Managed Service for Apache Flink に組み込まれている RANDOM_CUT_FOREST
関数を使用して、データストリームを処理し、データの異常を検出します。
B. リアルタイムデータを Amazon Kinesis Data Streams に取り込みます。Amazon SageMaker AI
エンドポイントをデプロイして、リアルタイムの外れ値検出を実現します。異常検出用の AWS Lambda 関数を作成します。データストリームを使用して Lambda
関数を呼び出します。
C. Amazon EC2 インスタンス上の Apache Kafka
にリアルタイムデータを取り込みます。Amazon SageMaker AI
エンドポイントをデプロイして、リアルタイムの外れ値検出を実現します。異常検出用の AWS Lambda 関数を作成します。データストリームを使用して Lambda
関数を呼び出します。
D. リアルタイムデータをAmazon Simple Queue Service (Amazon SQS) FIFOキューに送信します。キューメッセージを処理するためのAWS
Lambda関数を作成します。Lambda関数をプログラムして、AWS
Glueの抽出、変換、ロード (ETL) ジョブを開始し、バッチ処理と異常検出を行います。
Answer: A
Explanation:
The key requirements are real-time processing, high throughput, and minimal operational overhead. Amazon Kinesis Data Streams is designed for ingesting thousands of events per second with low latency.
For anomaly detection on streaming data, Amazon Managed Service for Apache Flink provides a built-in Random Cut Forest (RCF) function. RCF is an unsupervised anomaly detection algorithm that works well on numerical streaming data and does not require labeled training data.
This fully managed combination eliminates the need to deploy or maintain SageMaker endpoints, EC2 instances, or custom ML pipelines. Options B and C introduce unnecessary infrastructure and model management overhead. Option D is batch-oriented and unsuitable for real-time anomaly detection.
Therefore, using Kinesis Data Streams with Flink's built-in Random Cut Forest is the most scalable and low- overhead solution.

QUESTION NO: 12
MLエンジニアは、AWS Glue
DataBrewの最小最大正規化を用いてトレーニングデータを正規化しました。MLエンジニアは、本番環境の推論データをモデルに渡す前に、同様の方法で正規化する必要があります。
どのソリューションがこの要件を満たすでしょうか?
A. 既知のデータセットからの統計を適用して、生産サンプルを正規化します。

B. トレーニング
セットからの最小最大正規化統計を保持し、それを使用して製品サンプルを正規化します。
C.
一連の製造サンプルから新しい最小最大統計を計算し、それを使用してすべての製造サンプルを正規化します。
D.
各生産サンプルから新しい最小最大統計を計算し、それを使用してすべての生産サンプルを正規化します。
Answer: B
Explanation:
AWS ML best practices state that data preprocessing applied during training must be applied identically during inference. For min-max normalization, this requires reusing the minimum and maximum values calculated from the training dataset.
If production data is normalized using different statistics, the feature distributions will differ from what the model learned, leading to degraded prediction accuracy. AWS documentation explicitly warns against recomputing normalization parameters on inference data.
Options A, C, and D introduce data leakage or inconsistent feature scaling. Option B ensures consistency between training and inference pipelines and preserves model integrity.
Therefore, Option B is the correct and AWS-aligned solution.

QUESTION NO: 13
ケーススタディ
機械学習エンジニアがAWS上で不正検出モデルを開発しています。トレーニングデータセットには、オンプレミスのMySQLデータベースのトランザクションログ、顧客プロファイル、テーブルが含まれています。トランザクションログと顧客プロファイルはAmazon S3に保存されています。
データセットにはクラスの不均衡があり、モデルのアルゴリズムの学習に影響を与えています。さらに、多くの特徴量に相互依存性があり、アルゴリズムはデータ内の望ましい根本的なパターンをすべて捉えていません。
ML エンジニアがモデルをトレーニングする前に、ML
エンジニアは不均衡なデータの問題を解決する必要があります。
最も少ない運用労力でこの要件を満たすソリューションはどれでしょうか?
A. Amazon Athena
を使用して、不均衡の原因となるパターンを特定します。それに応じてデータセットを調整します。
B. Amazon SageMaker Studio Classic
の組み込みアルゴリズムを使用して、不均衡なデータセットを処理します。
C. AWS Glue DataBrew
の組み込み機能を使用して、少数クラスをオーバーサンプリングします。
D. Amazon SageMaker Data Wrangler
バランスデータ操作を使用して、少数クラスをオーバーサンプリングします。
Answer: D
Explanation:
Problem Description:
The training dataset has a class imbalance, meaning one class (e.g., fraudulent transactions)

has fewer samples compared to the majority class (e.g., non-fraudulent transactions). This imbalance affects the model's ability to learn patterns from the minority class.
Why SageMaker Data Wrangler?
SageMaker Data Wrangler provides a built-in operation called " Balance Data, " which includes oversampling and undersampling techniques to address class imbalances.
Oversampling the minority class replicates samples of the minority class, ensuring the algorithm receives balanced inputs without significant additional operational overhead.
Steps to Implement:
Import the dataset into SageMaker Data Wrangler.
Apply the " Balance Data " operation and configure it to oversample the minority class.
Export the balanced dataset for training.
Advantages:
Ease of Use: Minimal configuration is required.
Integrated Workflow: Works seamlessly with the SageMaker ecosystem for preprocessing and model training.
Time Efficiency: Reduces manual effort compared to external tools or scripts.

QUESTION NO: 14
機械学習エンジニアがシンプルなニューラルネットワークモデルを学習させています。検証データセットを用いて、モデルのパフォーマンスを経時的に追跡しています。モデルのパフォーマンスは当初大幅に向上しますが、一定数のエポックを過ぎると低下します。
この問題を緩和する解決策はどれですか? (2 つ選択してください。)
A. モデルの早期停止を有効にします。
B. レイヤー内のドロップアウトを増やします。
C. レイヤーの数を増やします。
D. ニューロンの数を増やします。
E. モデルのバイアスの原因を調査して削減します。
Answer: A B
Explanation:
Early stopping halts training once the performance on the validation dataset stops improving. This prevents the model from overfitting, which is likely the cause of performance degradation after a certain number of epochs.
Dropout is a regularization technique that randomly deactivates neurons during training, reducing overfitting by forcing the model to generalize better. Increasing dropout can help mitigate the problem of performance degradation due to overfitting.

QUESTION NO: 15
ある企業は、ユーザーのクリックに関する時系列データをAmazon S3バケットに保存しています。生データは、日々のユーザーアクティビティに関する数百万行のデータで構成されています。MLエンジニアは、このデータにアクセスしてMLモデルを開発しています。
MLエンジニアはAmazon Athenaを使用して毎日レポートを作成し、過去3日間のクリック傾向を分析する必要があります。会社はデータをアーカイブする前に30日間保持する必要があります。
どのソリューションがデータ取得に最高のパフォーマンスを提供しますか?

A.
時系列データはすべてS3バケットに分割せずに保存します。30日以上経過したデータは手動で別のS3バケットに移動します。
B. 時系列データを別々のS3バケットにコピーするためのAWS Lambda関数を作成します。S3ライフサイクルポリシーを適用し、30日以上経過したデータをS3 Glacier Flexible Retrievalにアーカイブします。
C.
時系列データをS3バケット内の日付プレフィックスごとにパーティションに整理します。S3ライフサイクルポリシーを適用し、30日以上経過したパーティションをS3 Glacier Flexible Retrievalにアーカイブします。
D.
各日の時系列データを専用のS3バケットに格納します。S3ライフサイクルポリシーを使用して、30日以上経過したデータを含むS3バケットをS3 Glacier Flexible Retrievalにアーカイブします。
Answer: C
Explanation:
Partitioning the time-series data by date prefix in the S3 bucket significantly improves query performance in Amazon Athena by reducing the amount of data that needs to be scanned during queries. This allows the ML engineers to efficiently analyze trends over specific time periods, such as the past 3 days. Applying S3 Lifecycle policies to archive partitions older than 30 days to S3 Glacier Flexible Retrieval ensures cost- effective data retention and storage management while maintaining high performance for recent data retrieval.

QUESTION NO: 16
カスタマーコールセンターは、Amazon Transcribeを使用して、顧客とサポート担当者間の数百件の音声録音をテキストファイルに変換しています。コールセンターは、これらのテキストファイルを使用して機械学習モデルをトレーニングしたいと考えています。業界規制を遵守するため、コールセンターはトレーニング用テキストファイルから顧客名、住所、電話番号を削除する必要があります。
開発労力を最小限に抑えつつ、これらの要件を満たすソリューションはどれでしょうか？
A. Amazon Bedrock Guardrails を使用して、テキスト ファイルから個人情報を処理および削除します。
B. AWS Glue Detect PII 変換を使用して、テキスト ファイルから個人情報を削除します。
C. テキストファイルをAmazon S3バケットに保存します。S3オブジェクトLambda関数を使用して個人情報を削除します。
D. テキストファイルから個人情報を削除するために、Amazon SageMaker Data Wrangler のカスタム変換を設定します。
Answer: B
Explanation:
Option B is correct because AWS Glue provides a built-in Detect PII transform that can detect, mask, or remove personally identifiable information with minimal custom development. AWS documentation says the Detect PII transform can process predefined AWS-managed PII entity types and supports actions such as removing or masking values. The examples in AWS docs explicitly mention sensitive entities such as phone numbers and

addresses, which directly match the problem statement.
The question specifically asks for the least development effort. That wording makes AWS Glue Detect PII the strongest answer because it is a native transformation capability rather than a custom code-heavy workflow.
AWS also documents fine-grained sensitive data detection features that let you apply actions per entity type, improving usability and reducing the need to build custom parsing and redaction logic yourself. This is much easier than creating Lambda-based transformation code or custom text-cleaning logic inside another ML preprocessing tool.
The other options are less suitable. Amazon Bedrock Guardrails is not the standard AWS service documented for bulk ETL-style redaction of training text files in this context. S3 Object Lambda would require more custom engineering to inspect and redact each object. SageMaker Data Wrangler custom transformation would also involve extra implementation work compared with using a purpose-built Glue transform. Because the call center already has text output and simply needs regulated fields like names, addresses, and phone numbers removed before training, the AWS-native low-effort solution is AWS Glue Detect PII. Therefore, the best verified answer is B.

QUESTION NO: 17
ある企業は顧客データを毎日収集し、日付ごとに分割されたAmazon S3バケットに圧縮ファイルとして保存しています。アナリストは毎月、データを処理し、データ品質をチェックし、結果をAmazon QuickSightダッシュボードにアップロードします。ML エンジニアは、データが QuickSight に送信される前に、最小限の運用オーバーヘッドでデータの品質を自動的にチェックする必要があります。
これらの要件を満たすソリューションはどれでしょうか?
A. AWS Glue クローラーを毎月実行し、AWS Glue データ品質ルールを使用してデータ品質をチェックします。
B. AWS Glue クローラーを実行し、PySpark を使用してカスタム AWS Glue ジョブを作成し、データ品質を評価します。
C. S3 アップロードによってトリガーされる Python スクリプトで AWS Lambda を使用して、データ品質を評価します。
D. S3 イベントを Amazon SQS に送信し、Amazon CloudWatch Insights を使用してデータ品質を評価します。
Answer: A
Explanation:
AWS Glue Data Quality provides managed, declarative data quality checks with minimal configuration.
Combined with Glue crawlers, it enables automatic schema discovery and quality validation without custom code.
Option A uses native AWS services designed for this exact purpose, minimizing operational overhead.
Options B and C require custom code and maintenance. Option D is not designed for data validation.
AWS documentation explicitly recommends Glue Data Quality rules for scalable, automated data quality checks in analytics pipelines.

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

QUESTION NO: 18
事例研究
ある企業がAmazon
SageMakerを使用して、ウェブベースのAIアプリケーションを構築しています。このアプリケーションは、機械学習の実験、トレーニング、中央モデルレジストリ、モデルのデプロイ、およびモデルの監視といった機能を提供します。
アプリケーションは、機械学習ライフサイクル全体を通して、トレーニングデータの安全かつ隔離された使用を保証する必要があります。トレーニングデータはAmazon
S3に保存されます。
同社は連続研修制度を試験的に導入している。
企業はこれらの業務におけるインフラの立ち上げ時間をどのように最小限に抑えることができるでしょうか？
A.マネージドスポットトレーニングを使用します。
B. SageMaker が管理するウォームプールを使用します。
C. SageMaker トレーニング コンパイラを使用します。
D. SageMaker分散データ並列処理(SMDDP)ライブラリを使用します。
Answer: B
Explanation:
When running consecutive training jobs in Amazon SageMaker, infrastructure provisioning can introduce latency, as each job typically requires the allocation and setup of compute resources. To minimize this startup time and enhance efficiency, Amazon SageMaker 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

QUESTION NO: 19
ある建設会社は、Amazon SageMaker AI
を使用して、道路の損傷を特定するためのカスタム物体検出モデルをトレーニングしています。複数のカメラから取得した画像を使用しています。画像は JPEG オブジェクトとして Amazon S3 バケットに保存されています。
画像は、トレーニングジョブで使用する前に、計算負荷の高いコンピュータービジョン技術を用いて前処理する必要があります。企業は、トレーニングジョブにおけるデータの読み込みと前処理を最適化する必要があります。このソリューションは、モデルのパフォーマンスに影響を与えたり、コンピューティングリソースやストレージリソースを増加したりしてはなりません。
これらの要件を満たすソリューションはどれでしょうか?
A. SageMaker AI ファイル モードを使用して、画像をバッチで読み込んで処理します。
B. モデルのバッチ サイズを縮小し、前処理スレッドの数を増やします。
C. S3 バケット内のトレーニング イメージの品質を下げます。
D. 画像を RecordIO 形式に変換し、遅延読み込みパターンを使用します。
Answer: D
Explanation:
AWS documentation recommends using RecordIO format with lazy loading to optimize data input pipelines for image-based training workloads. RecordIO is a binary data format that enables sequential reads, reducing I
/O overhead and improving throughput during training.
By converting JPEG images into RecordIO format, the training job can read data more efficiently from Amazon S3. Lazy loading ensures that only the required data is loaded into memory when needed, which optimizes CPU utilization during computationally intensive preprocessing steps.
Option A (file mode) results in many small S3 GET requests, which can become a bottleneck for large image datasets. Option B changes training behavior and can negatively affect convergence and performance. Option C reduces image quality, which directly impacts model accuracy and violates the requirement.
AWS SageMaker documentation explicitly highlights RecordIO and lazy loading as best practices for high- performance image training pipelines, especially when preprocessing is CPU-intensive.
Therefore, Option D is the correct and AWS-aligned solution.

QUESTION NO: 20

ある企業の機械学習エンジニアが、感情分析用の機械学習モデルをAmazon SageMakerエンドポイントにデプロイしました。この機械学習エンジニアは、モデルがどのように予測を行うのかを社内の関係者に説明する必要があります。
どのソリューションがモデルの予測を説明するでしょうか?
A. デプロイされたモデルで SageMaker Model Monitor を使用します。
B. デプロイされたモデルで SageMaker Clarify を使用します。
C. Amazon CloudWatch で A/# テストからの推論の分布を表示します。
D. シャドウエンドポイントを追加します。サンプルの予測の違いを分析します。
Answer: B
Explanation:
SageMaker Clarify is designed to provide explainability for ML models. It can analyze feature importance and explain how input features influence the model ' s predictions. By using Clarify with the deployed SageMaker model, the ML engineer can generate insights and present them to stakeholders to explain the sentiment analysis predictions effectively.

QUESTION NO: 21
ML エンジニアは、サブスクリプション
サービスの顧客離脱を予測するためのロジスティック回帰モデルを構築しています。
データセットには、location と job_seniority_level という 2
つの文字列変数が含まれています。
location 変数には 3 つの異なる値があり、job_seniority_level 変数には 10
を超える異なる値があります。
ML エンジニアは変数の前処理を実行する必要があります。
どのソリューションがこの要件を満たすでしょうか?
A.
場所にトークン化を適用します。job_seniority_levelに序数エンコーディングを適用します。
B.
locationにワンホットエンコーディングを適用します。job_seniority_levelに序数エンコーディングを適用します。
C. 場所にビニングを適用します。job_seniority_levelに標準スケーリングを適用します。
D.
locationにワンホットエンコーディングを適用します。job_seniority_levelに標準スケーリングを適用します。
Answer: B
Explanation:
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.

QUESTION NO: 22
ある企業は、Amazon SageMaker AI を使用して、CPU
上で実行されるリアルタイム予測用の機械学習モデルをホストしたいと考えています。このモデルは、営業時間中は断続的にトラフィックが発生し、営業時間後はトラフィックがゼロになる期間があります。
最もコスト効率の高い方法で推論リクエストを処理するホスティング
オプションはどれですか?
A. スケジュールされた自動スケーリングを使用して、モデルをリアルタイム
エンドポイントにデプロイします。
B. 営業時間中に、プロビジョニングされた同時実行性を使用して、SageMaker AI
Serverless Inference エンドポイントにモデルをデプロイします。
C. 自動スケーリングをゼロにして、モデルを非同期推論エンドポイントにデプロイします。
D. モデルをリアルタイムエンドポイントにデプロイし、AWS Lambda
を使用して営業時間内にのみアクティブ化します。
Answer: B
Explanation:
AWS recommends SageMaker Serverless Inference for workloads with intermittent or unpredictable traffic.
Serverless inference automatically scales compute resources to zero when idle, eliminating costs during periods with no traffic.
For business-hour traffic spikes, provisioned concurrency ensures low-latency responses while still avoiding the cost of continuously running instances. This model is especially cost-effective for CPU-based inference workloads.
Real-time endpoints incur costs even when idle, and asynchronous inference is designed for long-running jobs rather than low-latency predictions.
AWS documentation explicitly states that Serverless Inference is the most cost-effective option for intermittent real-time workloads.
Therefore, Option B is the correct choice.

QUESTION NO: 23
MLエンジニアは、遺伝的アルゴリズムに基づいてトレーニング済みのモデルをデプロイする必要があります。予測には数分かかる場合があり、リクエストには最大100MBのデータが含まれることがあります。
最も少ない運用オーバーヘッドでこれらの要件を満たす導入ソリューションはどれでしょう

か?
A. ALB の背後にある EC2 Auto Scaling にデプロイします。
B. SageMaker AI リアルタイムエンドポイントにデプロイします。
C. SageMaker AI 非同期推論エンドポイントにデプロイします。
D. EC2 上の Amazon ECS にデプロイします。
Answer: C
Explanation:
SageMaker Asynchronous Inference is designed for long-running inference workloads and large payloads (up to 1 GB). Requests are queued, processed asynchronously, and results are written to Amazon S3.
Real-time endpoints have payload and timeout limits. EC2 and ECS require infrastructure management, increasing operational overhead.
AWS documentation explicitly recommends asynchronous inference for workloads with large inputs and long execution times.
Therefore, Option C is the correct and most efficient solution.

QUESTION NO: 24
MLエンジニアがAmazon SageMaker AIでMLモデルを構築しています。MLエンジニアは、Amazon S3、Amazon Athena、Snowflakeから履歴データを直接SageMaker AIにロードする必要があります。どのソリューションがこの要件を満たすでしょうか?
A. AWS Glue DataBrew を使用してデータを SageMaker AI にインポートします。
B. SageMaker Pipelines でパイプラインを構築し、データを処理します。AWS DataSync を使用して、処理済みのデータを SageMaker AI にロードします。
C. SageMaker Feature Store に Feature Store を作成します。Apache Spark コネクタを使用して Feature Store に接続し、データにアクセスします。
D. SageMaker Data Wrangler を使用してデータをクエリおよびインポートします。
Answer: D
Explanation:
AWS provides Amazon SageMaker Data Wrangler as a native tool for importing, transforming, and analyzing data from multiple sources directly into SageMaker Studio. Data Wrangler supports Amazon S3, Amazon Athena, and Snowflake as built-in data sources through managed connectors.
Using Data Wrangler, ML engineers can query data from Athena using SQL, load structured files from S3, and securely connect to Snowflake without writing custom ingestion code. This approach significantly reduces development effort and aligns with AWS best practices for rapid ML experimentation.
Option A is incorrect because AWS Glue DataBrew is designed for data preparation but does not natively integrate with SageMaker training workflows. Option B introduces unnecessary complexity and is not intended for direct ML data loading. Option C focuses on feature storage, not raw historical data ingestion.
Therefore, SageMaker Data Wrangler is the correct solution.