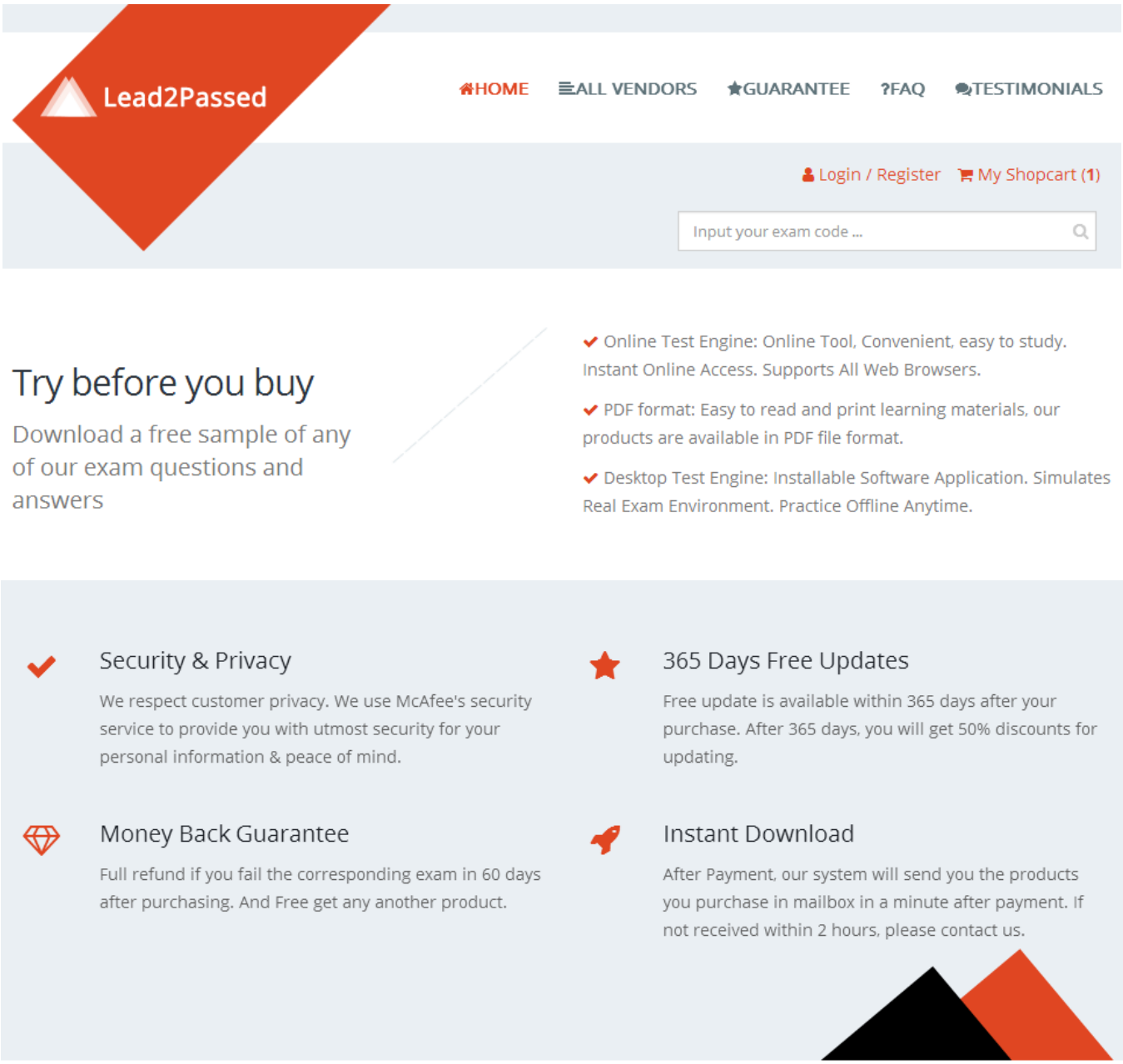
Lead2Passed



http://www.lead2passed.com

Valid Certification Exam Dumps Materials and Study Guide - Lead2Passed

Exam : **MLA-C01-KR**

Title : AWS Certified Machine Learning Engineer - Associate (MLA-C01 Korean Version)

Vendor : Amazon

Version : DEMO

QUESTION NO: 1
머신러닝 엔지니어가 Amazon SageMaker AI에서 학습 작업을 실행하려고 합니다. 이 학습 작업은 여러 개의 GPU를 사용하여 신경망을 학습시킵니다. 학습 데이터 세트는 Parquet 형식으로 저장되어 있습니다.
머신러닝 엔지니어는 Parquet 데이터셋에 SageMaker AI 학습 인스턴스의 메모리에 담기에는 너무 큰 파일들이 포함되어 있다는 사실을 발견했습니다.
어떤 해결책이 메모리 문제를 해결할까요?

A. Amazon Elastic Block Store(Amazon EBS) 프로비저닝된 IOPS SSD 볼륨을 인스턴스에 연결합니다.
파일을 EBS 볼륨에 저장하세요.

B. Amazon EMR의 Apache Spark를 사용하여 Parquet 파일을 재분할합니다. 재분할된 파일을 학습 작업에 사용합니다.

C. 학습 작업에 필요한 충분한 메모리를 갖춘 메모리 최적화 인스턴스로 인스턴스 유형을 변경합니다.

D. 메모리 사용량을 분산하려면 SageMaker AI 분산 데이터 병렬 처리(SMDDP) 라이브러리를 여러 인스턴스와 함께 사용하십시오.

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
머신러닝 엔지니어는 분석을 위해 아마존 S3에서 데이터를 사용하고, 준비하고, 로드하려고 합니다. 이를 위해서는 데이터의 스키마를 파악하고 메타데이터를 저장하기 위해 ETL(추출, 변환, 로드) 작업을 실행해야 합니다.
어떤 솔루션이 최소한의 수작업으로 이러한 요구 사항을 충족할까요?

A. AWS Glue를 사용하여 ETL 작업을 실행합니다. 해당 작업을 사용하여 스키마를 검색하고 관련 메타데이터를 AWS Glue 데이터 카탈로그에 저장합니다.

B. Amazon SageMaker Data Wrangler 플로우를 생성하여 ETL 작업을 실행합니다. 해당 작업을 사용하여 스키마를 검색하고 관련 메타데이터를 S3 버킷에 저장합니다.

C. Amazon Athena와 AWS Step Functions를 통합하여 ETL 파이프라인을 생성합니다. 이 파이프라인을 사용하여 ETL 작업을 실행하여 스키마를 검색하고 관련 메타데이터를 S3 버킷에 저장합니다.

D. scikit-learn 라이브러리가 포함된 Amazon EC2 인스턴스를 시작하여 ETL 작업을 실행합니다. 해당 작업을 사용하여 스키마를 검색하고 관련 메타데이터를 Amazon Redshift에 저장합니다.

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

한 회사에 벡터 데이터베이스를 사용하여 문서 임베딩을 저장하는 검색 증강 생성(RAG) 애플리케이션이 있습니다. 이 회사는 애플리케이션을 AWS로 마이그레이션하고 텍스트 파일에 대한 의미론적 검색을 제공하는 솔루션을 구현해야 합니다. 회사는 이미 텍스트 저장소를 Amazon S3 버킷으로 마이그레이션했습니다.

어떤 솔루션이 이러한 요구 사항을 충족시킬까요?

A. AWS Batch 작업을 사용하여 파일을 처리하고 임베딩을 생성합니다. AWS Glue를 사용하여 임베딩을 저장합니다. SQL 쿼리를 사용하여 의미 검색을 수행합니다.

B. 사용자 지정 Amazon SageMaker 노트북을 사용하여 사용자 지정 스크립트를 실행하여 임베딩을 생성합니다. SageMaker Feature Store를 사용하여 임베딩을 저장합니다. SQL 쿼리를 사용하여 의미 검색을 수행합니다.

C. Amazon Kendra S3 커넥터를 사용하여 S3 버킷에서 Amazon Kendra로 문서를 수집합니다. Amazon Kendra에 쿼리하여 의미 검색을 수행합니다.

D. Amazon Textract 비동기 작업을 사용하여 S3 버킷에서 문서를 수집합니다. Amazon Textract를 쿼리하여 의미 검색을 수행합니다.

Answer: C

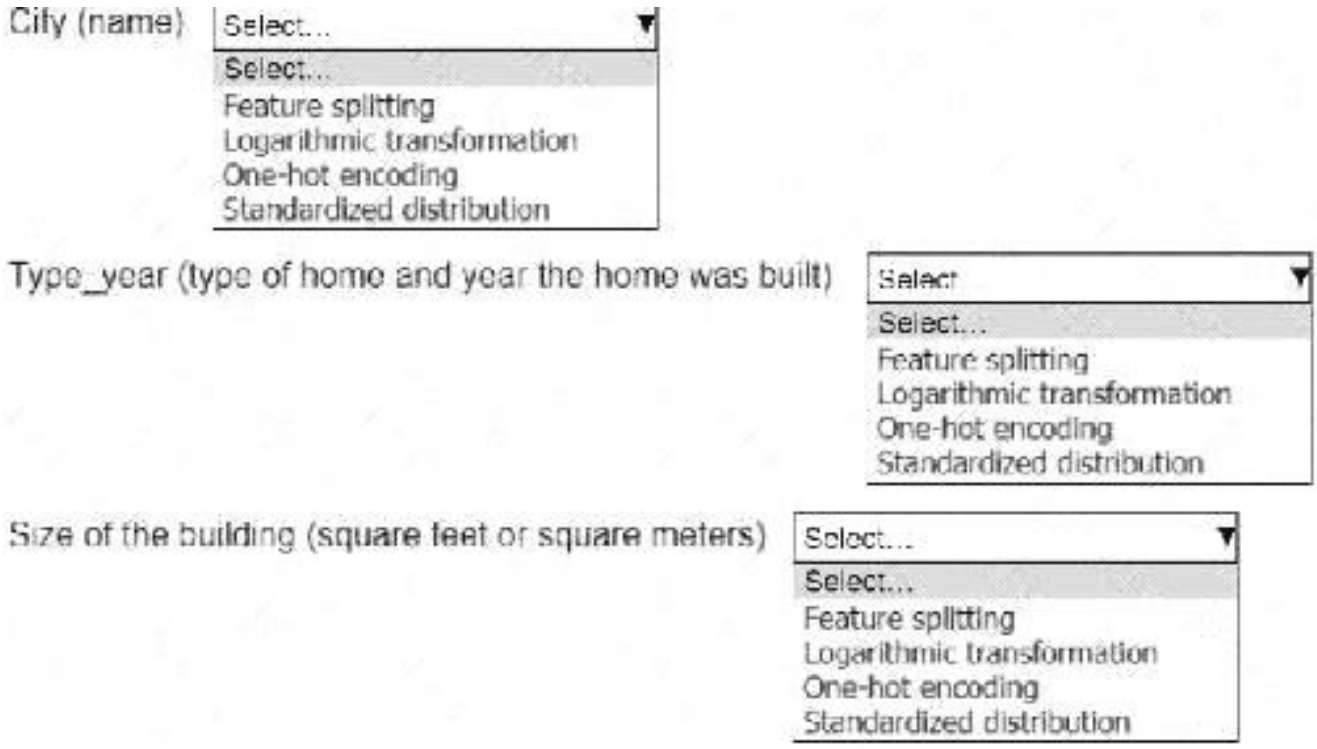
Explanation:

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

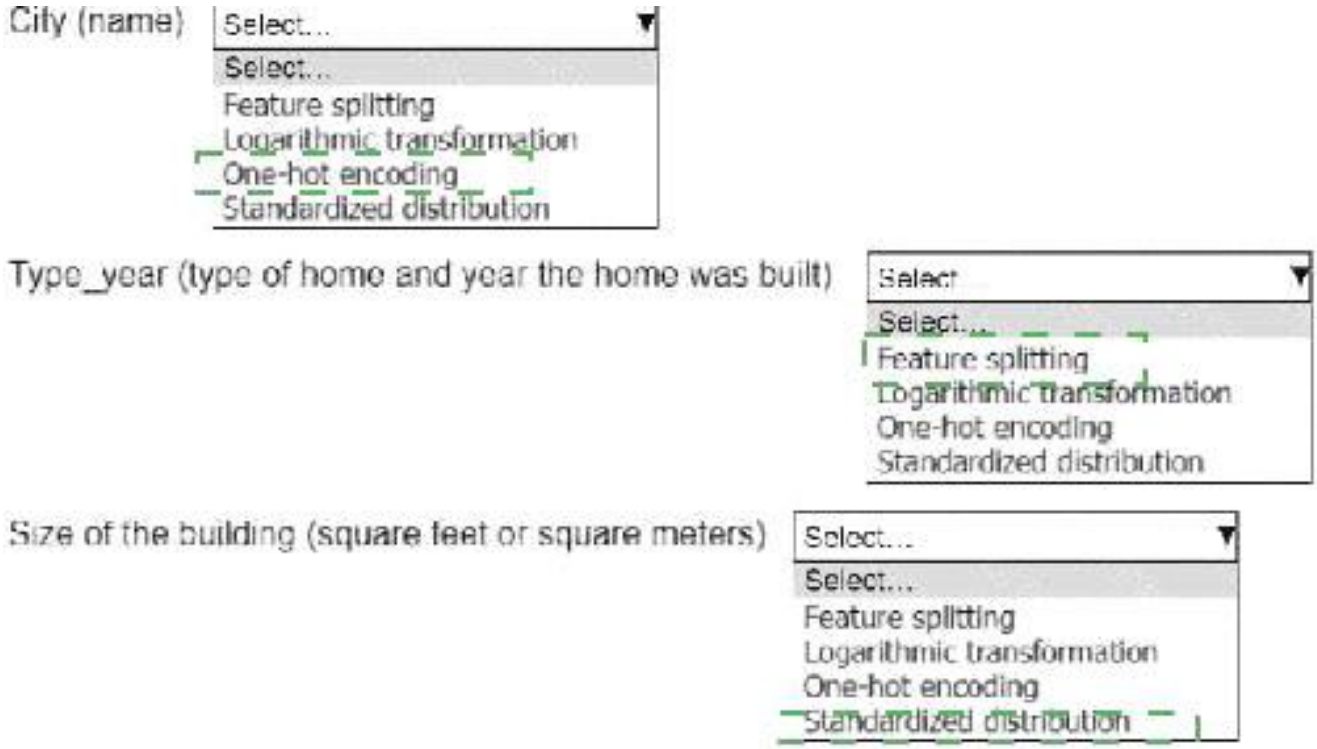
connector. Once the documents are ingested, Kendra enables semantic searches with its built-in capabilities, removing the need to manually generate embeddings or manage a vector database. This approach is efficient, requires minimal operational effort, and meets the requirements for a Retrieval Augmented Generation (RAG) application.

QUESTION NO: 4
한 회사가 Amazon SageMaker에서 ML 모델을 학습시켰습니다. 이 회사는 프로덕션 환경에서 추론을 제공하기 위해 해당 모델을 호스팅해야 합니다.
모델은 고가용성을 유지해야 하며 최소 지연 시간으로 응답해야 합니다. 각 요청의 크기는 1KB에서 3MB 사이입니다. 모델은 하루 중 예측할 수 없는 갑작스러운 요청을 받게 됩니다.
추론은 수요 변화에 비례하여 조정되어야 합니다.
이러한 요구 사항을 충족하기 위해 회사는 어떻게 모델을 프로덕션에 배포해야 할까요?
A. SageMaker 실시간 추론 엔드포인트를 생성합니다. 자동 크기 조정을 구성합니다. 기존 모델을 표시하도록 엔드포인트를 구성합니다.
B. Amazon Elastic Container Service(Amazon ECS) 클러스터에 모델을 배포합니다. ECS 클러스터의 CPU를 기반으로 하는 ECS 예약 확장을 사용합니다.
C. Amazon Elastic Kubernetes Service(Amazon EKS) 클러스터에 SageMaker Operator를 설치합니다. Amazon EKS에 모델을 배포합니다. 메모리 메트릭을 기준으로 복제본을 확장하도록 수평적 Pod 자동 확장을 설정합니다.
D. 추론을 위해 애플리케이션 로드 밸런서(ALB) 뒤에 스팟 플릿이 있는 스팟 인스턴스를 사용하세요. 자동 확장을 위한 지표로 ALBRequestCountPerTarget 지표를 사용하세요.
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
머신러닝 엔지니어가 비슷한 크기의 주택 가격을 예측하는 머신러닝 모델을 개발하고 있습니다. 이 모델은 여러 특징을 기반으로 예측을 수행할 것입니다. 머신러닝 엔지니어는 주택 가격을 예측하기 위해 다음과 같은 특징 엔지니어링 기법을 사용할 것입니다.
* 기능 분할
* 로그 변환
* 원핫 인코딩
* 표준화된 배포
다음 특징 목록에 맞는 특징 엔지니어링 기법을 선택하세요. 각 특징 엔지니어링 기법은 한 번만 선택하거나 전혀 선택하지 않아야 합니다. (세 가지를 선택하세요.)



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
한 회사는 ML 모델 공급업체로부터 정기적으로 새로운 학습 데이터를 받습니다. 공급업체는 3~4일마다 정리되고 준비된 데이터를 회사의 Amazon S3 버킷에 전달합니다.
이 회사는 모델을 재학습하기 위해 Amazon SageMaker 파이프라인을 사용하고 있습니다. ML 엔지니어는 S3 버킷에 새 데이터가 업로드될 때 파이프라인을 실행하는 솔루션을 구현해야 합니다.
어떤 솔루션이 최소한의 운영 노력으로 이러한 요구 사항을 충족할 수 있을까요?
A. SageMaker 학습 인스턴스로 데이터를 전송하고 학습을 시작하기 위한 S3 수명 주기 규칙을 만듭니다.
B. S3 버킷을 스캔하는 AWS Lambda 함수를 생성합니다. 새 데이터가 업로드되면 파이프라인을 시작하도록 Lambda 함수를 프로그래밍합니다.
C. S3 업로드와 일치하는 이벤트 패턴을 갖는 Amazon EventBridge 규칙을 생성합니다. 파이프라인을 규칙의 대상으로 구성합니다.
D. 새로운 데이터가 업로드될 때 파이프라인을 조정하려면 Apache Airflow(Amazon MWAA)용 Amazon Managed Workflows를 사용합니다.
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
한 회사가 머신러닝 모델의 데이터 수집 프로세스를 실행하기 위해 아마존 EMR 클러스터를 사용하고 있습니다. 머신러닝 엔지니어는 처리 시간이 증가하고 있음을 발견했습니다.
어떤 솔루션이 처리 시간을 가장 비용 효율적으로 단축할까요?
A. 스팟 인스턴스를 사용하여 기본 노드 수를 늘립니다.
B. 스팟 인스턴스를 사용하여 코어 노드 수를 늘립니다.
C. 스팟 인스턴스를 사용하여 작업 노드 수를 늘릴 수 있습니다.
D. 온디맨드 인스턴스를 사용하여 코어 노드 수를 늘립니다.
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
한 ML 엔지니어가 Amazon SageMaker에서 ML 모델을 학습시켜 회로 TV 영상에서 자동차 사고를 탐지하도록 했습니다. ML 엔지니어는 SageMaker Data Wrangler를 사용하여 사고 및 비사고 이미지로 구성된 학습 데이터 세트를 생성했습니다.
모델은 학습 및 검증 단계에서는 좋은 성능을 보였습니다. 그러나 다양한 카메라의 이미지 품질 차이로 인해 실제 운영 환경에서는 성능이 저하되었습니다.
어떤 솔루션이 가장 짧은 시간 안에 모델의 정확도를 향상시킬 수 있을까요?
A. 모든 카메라에서 더 많은 이미지를 수집합니다. Data Wrangler를 사용하여 새로운 학습 데이터 세트를 준비합니다.
B. Data Wrangler의 손상된 이미지 변환을 사용하여 학습 데이터 세트를 재생성합니다. 임펄스 노이즈 옵션을 지정합니다.
C. Data Wrangler의 이미지 대비 향상 변환을 사용하여 학습 데이터 세트를 재생성합니다. 감마 대비 옵션을 지정합니다.
D. Data Wrangler의 이미지 크기 조정 변환을 사용하여 학습 데이터 세트를 재생성합니다. 모든 이미지를 동일한 크기로 자릅니다.
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
머신러닝 엔지니어가 두 클래스 중 하나에서 성능이 저조한 이미지 분류 모델을 튜닝하고 있습니다. 성능이 저조한 클래스는 훈련 데이터 세트에서 극히 작은 부분을 차지합니다.
어떤 해결책이 모델의 성능을 향상시킬까요?
A. 정확도를 최적화합니다. 사용 빈도가 낮은 이미지에는 이미지 증강을 사용합니다.

B. F1 점수를 최적화합니다. 사용 빈도가 낮은 이미지에는 이미지 증강을 적용합니다.
C. 정확도 최적화. SMOTE를 사용하여 합성 이미지를 생성합니다.
D. F1 점수 최적화. SMOTE를 사용하여 합성 이미지 생성.
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
머신러닝 엔지니어가 주택 및 아파트 가격을 예측하는 모델을 구축하고 있습니다. 이 모델은 세 가지 특징을 사용합니다.
면적(제곱미터), 가격, 건물 연식에 대한 데이터 세트입니다. 총 10,000개의 행으로 구성되어 있으며, 대형 저택 하나와 초소형 아파트 하나에 대한 데이터가 포함되어 있습니다.
머신러닝 엔지니어는 모델이 일반적인 주택이나 아파트에 대해 정확한 예측을 생성할 수 있도록 데이터 세트에 대한 전처리 작업을 수행해야 합니다.
어떤 솔루션이 이러한 요구 사항을 충족할까요?
A. 이상치를 제거하고 제곱미터 변수에 로그 변환을 수행합니다.
B. 이상치를 유지하고 제곱미터 변수에 대해 정규화를 수행합니다.
C. 이상치를 제거하고 제곱미터 변수에 대해 원핫 인코딩을 수행합니다.
D. 이상치를 유지하고 제곱미터 변수에 대해 원핫 인코딩을 수행합니다.
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
한 금융 회사가 외부 공급업체로부터 대량의 실시간 시장 데이터 스트림을 수신합니다. 이 스트림은 매초 수천 개의 JSON 레코드로 구성됩니다.
해당 회사는 이상 데이터 포인트를 식별하기 위해 AWS에 확장 가능한 솔루션을 구현해야 합니다.
어떤 솔루션이 운영 부담을 최소화하면서 이러한 요구 사항을 충족할까요?
A. 실시간 데이터를 Amazon Kinesis Data Streams로 수집합니다. Amazon Managed Service for Apache Flink에 내장된 RANDOM_CUT_FOREST 함수를 사용하여 데이터 스트림을 처리하고 데이터 이상을 감지합니다.
B. 실시간 데이터를 Amazon Kinesis Data Streams로 수집합니다. 실시간 이상치 감지를 위해 Amazon SageMaker AI 엔드포인트를 배포합니다. 이상치를 감지하는 AWS Lambda 함수를 생성합니다. 데이터 스트림을 사용하여 Lambda 함수를 호출합니다.
C. Amazon EC2 인스턴스의 Apache Kafka에 실시간 데이터를 수집합니다. 실시간 이상치 탐지를 위해 Amazon SageMaker AI 엔드포인트를 배포합니다. 이상치를 탐지하는 AWS Lambda 함수를 생성합니다. 데이터 스트림을 사용하여 Lambda 함수를 호출합니다.
D. 실시간 데이터를 Amazon Simple Queue Service(Amazon SQS) FIFO 큐로 전송합니다. 큐 메시지를 처리하는 AWS Lambda 함수를 생성합니다. Lambda 함수가 배치 처리 및 이상 탐지를 위해 AWS Glue ETL(추출, 변환 및 로드) 작업을 시작하도록 프로그래밍합니다.
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
머신러닝 엔지니어는 AWS Glue DataBrew에서 최소-최대 정규화를 사용하여 학습 데이터를 정규화했습니다. 이제 머신러닝 엔지니어는 프로덕션 추론 데이터를 모델에 전달하기 전에 동일한 방식으로 정규화해야 합니다.
어떤 솔루션이 이 요구 사항을 충족할까요?

A. 잘 알려진 데이터 세트의 통계를 적용하여 프로덕션 샘플을 정규화합니다.
B. 훈련 세트의 최소-최대 정규화 통계를 유지하고 이를 사용하여 프로덕션 샘플을 정규화합니다.
C. 생산 샘플 배치에서 새로운 최소-최대 통계를 계산하고 이를 사용하여 모든 생산 샘플을 정규화합니다.
D. 각 생산 샘플에서 새로운 최소-최대 통계를 계산하고 이를 사용하여 모든 생산 샘플을 정규화합니다.
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
사례 연구
머신러닝 엔지니어가 AWS에서 사기 탐지 모델을 개발하고 있습니다. 학습 데이터 세트에는 온프레미스 MySQL 데이터베이스의 거래 로그, 고객 프로필, 테이블이 포함됩니다. 거래 로그와 고객 프로필은 Amazon S3에 저장됩니다.
데이터셋에는 모델 알고리즘의 학습에 영향을 미치는 클래스 불균형이 있습니다. 또한, 많은 특성들이 상호 의존성을 가지고 있습니다. 알고리즘이 데이터에서 원하는 모든 기본 패턴을 포착하지 못하고 있습니다.
ML 엔지니어가 모델을 학습시키기 전에 ML 엔지니어는 불균형한 데이터 문제를 해결해야 합니다.
어떤 솔루션이 최소한의 운영 노력으로 이 요구 사항을 충족할 수 있을까요?
A. Amazon Athena를 사용하여 불균형에 영향을 미치는 패턴을 파악합니다. 이에 따라 데이터 세트를 조정합니다.
B. Amazon SageMaker Studio Classic의 기본 제공 알고리즘을 사용하여 불균형 데이터 세트를 처리합니다.
C. AWS Glue DataBrew의 기본 제공 기능을 사용하여 소수 계층을 과도하게 샘플링합니다.
D. Amazon SageMaker Data Wrangler 균형 데이터 작업을 사용하여 소수 계층을 과도하게 샘플링합니다.
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
ML 엔지니어가 간단한 신경망 모델을 학습하고 있습니다. ML 엔지니어는 검증 데이터셋에서 시간 경과에 따른 모델의 성능을 추적합니다. 모델의 성능은 처음에는 상당히 향상되다가 특정 에포크(epoch) 수를 지나면서 저하됩니다.
어떤 해결책이 이 문제를 완화할 수 있을까요? (두 가지를 선택하세요.)
A. 모델에서 조기 중지를 활성화합니다.
B. 레이어의 드롭아웃을 증가시킵니다.
C. 레이어의 수를 늘립니다.
D. 뉴런의 수를 늘립니다.
E. 모델 편향의 원인을 조사하고 줄입니다.
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
한 회사가 사용자 클릭에 대한 시계열 데이터를 Amazon S3 버킷에 저장합니다. 원시 데이터는 매일 수백만 행의 사용자 활동으로 구성됩니다. ML 엔지니어는 이 데이터에 액세스하여 ML 모델을 개발합니다.
ML 엔지니어는 Amazon Athena를 사용하여 일일 보고서를 생성하고 지난 3일간의 클릭 추세를 분석해야 합니다. 회사는 데이터를 보관하기 전에 30일 동안 데이터를 보관해야 합니다.
어떤 솔루션이 데이터 검색에서 가장 높은 성능을 제공할까요?
A. 모든 시계열 데이터를 분할하지 않고 S3 버킷에 보관합니다. 30일이 지난 데이터는 수동으로 별도의 S3 버킷으로 이동합니다.
B. AWS Lambda 함수를 생성하여 시계열 데이터를 별도의 S3 버킷에 복사합니다. S3 수명 주기 정책을 적용하여 30일 이상 경과된 데이터를 S3 Glacier Flexible Retrieval에 보관합니다.

C. S3 버킷에서 시계열 데이터를 날짜 접두사별로 파티션으로 구성합니다. 30일 이상 경과된 파티션은 S3 Glacier Flexible Retrieval에 보관하려면 S3 수명 주기 정책을 적용합니다.
D. 각 날짜의 시계열 데이터를 별도의 S3 버킷에 저장합니다. S3 수명 주기 정책을 사용하여 30일 이상 경과된 데이터가 있는 S3 버킷을 S3 Glacier Flexible Retrieval에 보관합니다.

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

고객 콜센터에서 아마존 트랜스크라이브를 사용하여 고객과 상담원 간의 대화가 녹음된 수백 개의 오디오 파일을 텍스트 파일로 변환합니다. 콜센터는 이 텍스트 파일을 머신러닝 모델 학습에 사용하려고 합니다. 업계 규정을 준수하기 위해 콜센터는 학습용 텍스트 파일에서 고객의 이름, 주소, 전화번호를 삭제해야 합니다.
어떤 솔루션이 최소한의 개발 노력으로 이러한 요구 사항을 충족할까요?

A. Amazon Bedrock Guardrails를 사용하여 텍스트 파일에서 개인 정보를 처리하고 수정합니다.
B. AWS Glue Detect PII 변환을 사용하여 텍스트 파일에서 개인 정보를 제거합니다.
C. 텍스트 파일을 Amazon S3 버킷에 저장합니다. S3 Object Lambda 함수를 사용하여 개인 정보를 가립니다.
D. 텍스트 파일에서 개인 정보를 제거하도록 Amazon SageMaker Data Wrangler 사용자 지정 변환을 구성합니다.

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
한 회사가 고객 데이터를 매일 수집하여 날짜별로 파티션된 압축 파일 형태로 Amazon S3 버킷에 저장합니다. 매달 분석가들은 데이터를 처리하고 데이터 품질을 확인한 후 결과를 Amazon QuickSight 대시보드에 업로드합니다.
머신러닝 엔지니어는 최소한의 운영 부담으로 데이터를 QuickSight로 전송하기 전에 데이터 품질을 자동으로 검사해야 합니다.
어떤 솔루션이 이러한 요구 사항을 충족할까요?
A. AWS Glue 크롤러를 매월 실행하고 AWS Glue 데이터 품질 규칙을 사용하여 데이터 품질을 확인합니다.
B. AWS Glue 크롤러를 실행하고 PySpark를 사용하여 사용자 지정 AWS Glue 작업을 생성하여 데이터 품질을 평가합니다.
C. S3 업로드에 의해 트리거되는 Python 스크립트와 AWS Lambda를 사용하여 데이터 품질을 평가합니다.
D. S3 이벤트를 Amazon SQS로 전송하고 Amazon CloudWatch Insights를 사용하여 데이터 품질을 평가합니다.
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
사례 연구
한 회사가 아마존 세이지메이커(Amazon SageMaker)를 사용하여 웹 기반 AI 애플리케이션을 개발하고 있습니다. 이 애플리케이션은 머신러닝 실험, 학습, 중앙 모델 등록, 모델 배포 및 모델 모니터링과 같은 기능을 제공할 예정입니다.
해당 애플리케이션은 머신러닝 개발 주기 동안 학습 데이터의 안전하고 격리된 사용을 보장해야 합니다. 학습 데이터는 Amazon S3에 저장됩니다.
회사는 연속 교육 직무를 시험적으로 도입하고 있습니다.
회사는 이러한 직무에 필요한 인프라 구축 시간을 어떻게 최소화할 수 있을까요?
A. 관리형 스팟 트레이닝을 사용합니다.
B. SageMaker에서 관리하는 웜 풀을 사용합니다.

C. SageMaker 교육용 컴파일러를 사용하세요.
D. SageMaker 분산 데이터 병렬 처리(SMDDP) 라이브러리를 사용합니다.
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
한 건설 회사가 아마존 세이지메이커 AI를 사용하여 도로 손상을 식별하는 특수 맞춤형 객체 감지 모델을 학습시키고 있습니다. 이 회사는 여러 대의 카메라에서 촬영한 이미지를 사용하며, 이 이미지들은 아마존 S3 버킷에 JPEG 객체로 저장됩니다.
이미지를 학습 작업에 사용하기 전에 연산 집약적인 컴퓨터 비전 기술을 사용하여 전처리해야 합니다. 회사는 학습 작업에서 데이터 로딩 및 전처리를 최적화해야 합니다. 솔루션은 모델

성능에 영향을 미치거나 컴퓨팅 또는 스토리지 리소스를 증가시켜서는 안 됩니다.
어떤 솔루션이 이러한 요구 사항을 충족할까요?
A. SageMaker AI 파일 모드를 사용하여 이미지를 일괄적으로 불러오고 처리합니다.
B. 모델의 배치 크기를 줄이고 전처리 스레드 수를 늘립니다.
C. S3 버킷에 있는 학습 이미지의 품질을 낮춥니다.
D. 이미지를 RecordIO 형식으로 변환하고 지연 로딩 패턴을 사용합니다.
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
한 회사의 ML 엔지니어가 Amazon SageMaker 엔드포인트에 감정 분석을 위한 ML 모델을 배포했습니다. ML 엔지니어는 회사 이해관계자들에게 모델의 예측 방식을 설명해야 합니다.
어떤 솔루션이 모델의 예측에 대한 설명을 제공할까요?
A. 배포된 모델에서 SageMaker Model Monitor를 사용합니다.
B. 배포된 모델에 SageMaker Clarify를 사용합니다.
C. Amazon CloudWatch에서 A/# 테스트의 추론 분포를 보여줍니다.
D. 그림자 엔드포인트를 추가합니다. 샘플의 예측 차이를 분석합니다.
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
머신러닝 엔지니어가 구독 서비스 고객 이탈을 예측하는 로지스틱 회귀 모델을 구축하고 있습니다.
데이터 세트에는 location과 job_seniority_level이라는 두 개의 문자열 변수가 포함되어 있습니다.

location 변수는 3개의 서로 다른 값을 가지며, job_seniority_level 변수는 10개 이상의 서로 다른 값을 가집니다.
머신러닝 엔지니어는 변수에 대한 전처리 작업을 수행해야 합니다.
어떤 솔루션이 이 요구 사항을 충족할까요?
A. 위치에 토큰화를 적용합니다. job_seniority_level에 서수 인코딩을 적용합니다.
B. 위치에 원핫 인코딩을 적용합니다. job_seniority_level에 서수 인코딩을 적용합니다.
C. 위치에 빈닝을 적용합니다. job_seniority_level에 표준 스케일링을 적용합니다.
D. 위치에 원핫 인코딩을 적용합니다. job_seniority_level에 표준 스케일링을 적용합니다.
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
한 회사가 실시간 예측을 위해 CPU에서 실행되는 머신러닝 모델을 호스팅하기 위해 Amazon SageMaker AI를 사용하려고 합니다. 이 모델은 업무 시간 동안에는 간헐적으로 트래픽이 발생하고, 업무 시간 이후에는 트래픽이 없는 기간이 있습니다.
어떤 호스팅 옵션이 추론 요청을 가장 비용 효율적으로 처리할까요?
A. 예약된 자동 확장 기능을 사용하여 모델을 실시간 엔드포인트에 배포합니다.
B. 업무 시간 동안 프로비저닝된 동시 실행 수로 모델을 SageMaker AI 서버리스 추론 엔드포인트에 배포합니다.
C. 자동 스케일링을 0으로 설정하여 모델을 비동기 추론 엔드포인트에 배포합니다.
D. AWS Lambda를 사용하여 모델을 실시간 엔드포인트에 배포하고 업무 시간 동안에만 활성화합니다.
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
머신러닝 엔지니어는 유전 알고리즘을 기반으로 학습된 모델을 배포해야 합니다. 예측에는 몇 분이 소요될 수 있으며, 요청에는 최대 100MB의 데이터가 포함될 수 있습니다.
어떤 배포 솔루션이 운영 부담을 최소화하면서 이러한 요구 사항을 충족할까요?
A. ALB 뒤의 EC2 자동 스케일링에 배포합니다.
B. SageMaker AI 실시간 엔드포인트에 배포합니다.
C. SageMaker AI 비동기 추론 엔드포인트에 배포합니다.
D. EC2의 Amazon ECS에 배포합니다.
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
머신러닝 엔지니어가 Amazon SageMaker AI에서 머신러닝 모델을 구축하고 있습니다. 이 엔지니어는 Amazon S3, Amazon Athena 및 Snowflake에서 과거 데이터를 SageMaker AI로 직접 로드해야 합니다.
어떤 솔루션이 이 요구 사항을 충족할까요?
A. AWS Glue DataBrew를 사용하여 데이터를 SageMaker AI로 가져옵니다.
B. SageMaker Pipelines에서 데이터 처리를 위한 파이프라인을 구축합니다. AWS DataSync를 사용하여 처리된 데이터를 SageMaker AI에 로드합니다.
C. SageMaker Feature Store에서 피처 스토어를 생성합니다. Apache Spark 커넥터를 사용하여 Feature Store의 데이터에 액세스합니다.
D. SageMaker Data Wrangler를 사용하여 데이터를 쿼리하고 가져옵니다.
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.