MLS-C01 Q&As

AWS Certified Machine Learning - Specialty (MLS-C01)

Pass Amazon MLS-C01 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.pass2lead.com/aws-certified-machine-learning-specialty.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Amazon Official Exam Center

- **Instant Download** After Purchase
- **100% Money Back** Guarantee
- **365 Days** Free Update
- **800,000+** Satisfied Customers

QUESTION 1

A data scientist at a financial services company used Amazon SageMaker to train and deploy a model that predicts loan defaults. The model analyzes new loan applications and predicts the risk of loan default. To train the model, the data scientist manually extracted loan data from a database. The data scientist performed the model training and deployment steps in a Jupyter notebook that is hosted on SageMaker Studio notebooks. The model\\'s prediction accuracy is decreasing over time.

Which combination of steps is the MOST operationally efficient way for the data scientist to maintain the model\\'s accuracy? (Choose two.)

A. Use SageMaker Pipelines to create an automated workflow that extracts fresh data, trains the model, and deploys a new version of the model.

B. Configure SageMaker Model Monitor with an accuracy threshold to check for model drift. Initiate an Amazon CloudWatch alarm when the threshold is exceeded. Connect the workflow in SageMaker Pipelines with the CloudWatch alarm to automatically initiate retraining.

C. Store the model predictions in Amazon S3. Create a daily SageMaker Processing job that reads the predictions from Amazon S3, checks for changes in model prediction accuracy, and sends an email notification if a significant change is detected.

D. Rerun the steps in the Jupyter notebook that is hosted on SageMaker Studio notebooks to retrain the model and redeploy a new version of the model.

E. Export the training and deployment code from the SageMaker Studio notebooks into a Python script. Package the script into an Amazon Elastic Container Service (Amazon ECS) task that an AWS Lambda function can initiate.

Correct Answer: AB

QUESTION 2

A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.

The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.

Which solution should the Data Scientist build to satisfy the requirements?

A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.

B. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.

C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL

database. Have the Analysts query and run dashboards from the RDS database.

D. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to Bl tools using the Athena Java Database Connectivity (JDBC) connector.

Correct Answer: A

Firehose does integrate with GLue data catalog and it also "Buffers" the data .

"When Kinesis Data Firehose processes incoming events and converts the data to Parquet, it needs to know which schema to apply." This is achived by glue data catalog and athena and it works on real-time data ingest.See link below.

https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/

QUESTION 3

A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:

1.

Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.

2.

Support event-driven ETL pipelines.

3.

Provide a quick and easy way to understand metadata. Which approach meets trfese requirements?

A. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Glue ETL job, and an AWS Glue Data catalog to search and discover metadata.

B. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Batch job, and an external Apache Hive metastore to search and discover metadata.

C. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Batch job, and an AWS Glue Data Catalog to search and discover metadata.

D. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Glue ETL job, and an external Apache Hive metastore to search and discover metadata.

Correct Answer: A

The AWS Glue Data Catalog is your persistent metadata store. It is a managed service that lets you store, annotate, and share metadata in the AWS Cloud in the same way you would in an Apache Hive metastore. The Data Catalog is a drop-in replacement for the Apache Hive Metastore

https://docs.aws.amazon.com/zh_tw/glue/latest/dg/components-overview.html

QUESTION 4

An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget.

What should the Specialist do to meet these requirements?

A. Create one-hot word encoding vectors.

B. Produce a set of synonyms for every word using Amazon Mechanical Turk.

C. Create word embedding factors that store edit distance with every other word.

D. Download word embedding\\'s pre-trained on a large corpus.

Correct Answer: D

Word embeddings are a type of dense representation of words, which encode semantic meaning in a vector form. These embeddings are typically pre-trained on a large corpus of text data, such as a large set of books, news articles, or web pages, and capture the context in which words are used. Word embeddings can be used as features for a nearest neighbor model, which can be used to find words used in similar contexts.

Downloading pre-trained word embeddings is a good way to get started quickly and leverage the strengths of these representations, which have been optimized on a large amount of data. This is likely to result in more accurate and reliable features than other options like one-hot encoding, edit distance, or using Amazon Mechanical Turk to produce synonyms.

QUESTION 5

A company has a podcast platform that has thousands of users. The company implemented an algorithm to detect low podcast engagement based on a 10-minute running window of user events such as listening to, pausing, and closing the podcast. A machine learning (ML) specialist is designing the ingestion process for these events. The ML specialist needs to transform the data to prepare the data for inference.

How should the ML specialist design the transformation step to meet these requirements with the LEAST operational effort?

A. Use an Amazon Managed Streaming for Apache Kafka (Amazon MSK) cluster to ingest event data. Use Amazon Kinesis Data Analytics to transform the most recent 10 minutes of data before inference.

B. Use Amazon Kinesis Data Streams to ingest event data. Store the data in Amazon S3 by using Amazon Kinesis Data Firehose. Use AWS Lambda to transform the most recent 10 minutes of data before inference.

C. Use Amazon Kinesis Data Streams to ingest event data. Use Amazon Kinesis Data Analytics to transform the most recent 10 minutes of data before inference.

D. Use an Amazon Managed Streaming for Apache Kafka (Amazon MSK) cluster to ingest event data. Use AWS Lambda to transform the most recent 10 minutes of data before inference.

Correct Answer: C

QUESTION 6

A company wants to conduct targeted marketing to sell solar panels to homeowners. The company wants to use machine learning (ML) technologies to identify which houses already have solar panels. The company has collected 8,000 satellite images as training data and will use Amazon SageMaker Ground Truth to label the data.

The company has a small internal team that is working on the project. The internal team has no ML expertise and no ML experience.

Which solution will meet these requirements with the LEAST amount of effort from the internal team?

A. Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Use Amazon Rekognition Custom Labels for model training and hosting.

B. Set up a private workforce that consists of the internal team. Use the private workforce to label the data. Use Amazon Rekognition Custom Labels for model training and hosting.

C. Set up a private workforce that consists of the internal team. Use the private workforce and the SageMaker Ground Truth active learning feature to label the data. Use the SageMaker Object Detection algorithm to train a model. Use SageMaker batch transform for inference.

D. Set up a public workforce. Use the public workforce to label the data. Use the SageMaker Object Detection algorithm to train a model. Use SageMaker batch transform for inference.

Correct Answer: B

QUESTION 7

An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time.

Which solution should the agency consider?

A. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees, and alert when non-employees are detected.

B. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Image to detect faces from a collection of known employees and alert when non-employees are detected.

C. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when nonemployees are detected.

D. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, run an AWS Lambda function to capture image fragments and then call Amazon Rekognition Image to detect faces from a collection of known employees, and alert when non-employees are detected.

Correct Answer: A

https://aws.amazon.com/rekognition/video-features/

QUESTION 8

A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena The dataset contains more than 800.000 records stored as plaintext CSV files Each record contains 200 columns and is approximately 1 5 MB in size Most queries will span 5 to 10 columns only

How should the Machine Learning Specialist transform the dataset to minimize query runtime?

A. Convert the records to Apache Parquet format

B. Convert the records to JSON format

C. Convert the records to GZIP CSV format

D. Convert the records to XML format

Correct Answer: A

Using compressions will reduce the amount of data scanned by Amazon Athena, and also reduce your S3 bucket storage. It\\'s a Win-Win for your AWS bill. Supported formats: GZIP, LZO, SNAPPY (Parquet) and ZLIB.

Reference: https://www.cloudforecast.io/blog/using-parquet-on-athena-to-save-money-on-aws/

QUESTION 9

A machine learning (ML) engineer is integrating a production model with a customer metadata repository for real-time inference. The repository is hosted in Amazon SageMaker Feature Store. The engineer wants to retrieve only the latest version of the customer metadata record for a single customer at a time.

Which solution will meet these requirements?

A. Use the SageMaker Feature Store BatchGetRecord API with the record identifier. Filter to find the latest record.

B. Create an Amazon Athena query to retrieve the data from the feature table.

C. Create an Amazon Athena query to retrieve the data from the feature table. Use the write_time value to find the latest record.

D. Use the SageMaker Feature Store GetRecord API with the record identifier.

Correct Answer: D

QUESTION 10

A Machine Learning Specialist is configuring automatic model tuning in Amazon SageMaker

When using the hyperparameter optimization feature, which of the following guidelines should be followed to improve optimization?

Choose the maximum number of hyperparameters supported by

A. Amazon SageMaker to search the largest number of combinations possible

B. Specify a very large hyperparameter range to allow Amazon SageMaker to cover every possible value.

C. Use log-scaled hyperparameters to allow the hyperparameter space to be searched as quickly as possible

D. Execute only one hyperparameter tuning job at a time and improve tuning through successive rounds of experiments

Correct Answer: C

QUESTION 11

A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data

available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.

Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations

appear normal compared to the rest of the sample population.

How should the Data Scientist correct this issue?

A. Drop all records from the dataset where age has been set to 0.

B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset.

C. Drop the age feature from the dataset and train the model using the rest of the features.

D. Use k-means clustering to handle missing features.

Correct Answer: D

Dropping the Age feature is a NOT ATOLL a good idea - as age plays a critical role in this disease as per the question Dropping 10% of data is NOT a good idea considering the fact that the number of observations is already low. The Mean or Median are a potential solutions But the question says that "Disease worsens after age 65 so there is a correlation between age and other symptoms related feature" So that means that using Unsupervised Learning we can make pretty good prediction of "Age" So the answer is D Use K-Means clustering

QUESTION 12

A sports analytics company is providing services at a marathon. Each runner in the marathon will have their race ID printed as text on the front of their shirt. The company needs to extract race IDs from images of the runners. Which solution will meet these requirements with the LEAST operational overhead?

A. Use Amazon Rekognition.

B. Use a custom convolutional neural network (CNN).

C. Use the Amazon SageMaker Object Detection algorithm.

D. Use Amazon Lookout for Vision.

Correct Answer: A

https://docs.aws.amazon.com/rekognition/latest/dg/text-detection.html?pg=lnandsec=ft

QUESTION 13

A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined The model needs lo be retrained daily

Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

A. Require that the stores to switch to capturing their data locally on AWS Storage Gateway for loading into Amazon S3 then use AWS Glue to do the transformation

B. Deploy an Amazon EMR cluster running Apache Spark with the transformation logic, and have the cluster run each day on the accumulating records in Amazon S3, outputting new/transformed records to Amazon S3

C. Spin up a fleet of Amazon EC2 instances with the transformation logic, have them transform the data records accumulating on Amazon S3, and output the transformed records to Amazon S3.

D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehouse stream that transforms raw record attributes into simple transformed values using SQL.

Correct Answer: D

https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html

QUESTION 14

An online advertising company is developing a linear model to predict the bid price of advertisements in real time with low-latency predictions. A data scientist has trained the linear model by using many features, but the model is overfitting the training dataset. The data scientist needs to prevent overfitting and must reduce the number of features.

Which solution will meet these requirements?

A. Retrain the model with L1 regularization applied.

B. Retrain the model with L2 regularization applied.

C. Retrain the model with dropout regularization applied.

D. Retrain the model by using more data.

Correct Answer: A

https://www.google.com/url?sa=tandrct=jandq=andesrc=sandsource=webandcd=andcad=rjaanduact=8andved=2ahukewim5ygd0-x9ahxpycakhsalakoqfnoeca0qaqandurl=https%3a%2f%2fdocs.aws.amazon.com%2fmachine-

learning%2flatest%2fdg%2fmodel-fitunderfitting-vs-overfitting.htmlandusg=aovvaw2jwlt-j0jrsweidyjezi_s

QUESTION 15

A data engineer is using AWS Glue to create optimized, secure datasets in Amazon S3. The data science team wants the ability to access the ETL scripts directly from Amazon SageMaker notebooks within a VPC. After this setup is complete, the data science team wants the ability to run the AWS Glue job and invoke the SageMaker training job.

Which combination of steps should the data engineer take to meet these requirements? (Choose three.)

A. Create a SageMaker development endpoint in the data science team\\'s VPC.

B. Create an AWS Glue development endpoint in the data science team\\'s VPC.

C. Create SageMaker notebooks by using the AWS Glue development endpoint.

D. Create SageMaker notebooks by using the SageMaker console.

E. Attach a decryption policy to the SageMaker notebooks.

F. Create an IAM policy and an IAM role for the SageMaker notebooks.

Correct Answer: BCF

https://docs.aws.amazon.com/glue/latest/dg/dev-endpoint-tutorial-sage.html

Latest MLS-C01 Dumps MLS-C01 PDF Dumps MLS-C01 Braindumps