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.passapply.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 Machine Learning Specialist has completed a proof of concept for a company using a small data sample and now the Specialist is ready to implement an end-to-end solution in AWS using Amazon SageMaker The historical training data is stored in Amazon RDS Which approach should the Specialist use for training a model using that data?

A. Write a direct connection to the SQL database within the notebook and pull data in

B. Push the data from Microsoft SQL Server to Amazon S3 using an AWS Data Pipeline and provide the S3 location within the notebook.

C. Move the data to Amazon DynamoDB and set up a connection to DynamoDB within the notebook to pull data in

D. Move the data to Amazon ElastiCache using AWS DMS and set up a connection within the notebook to pull data in for fast access.

Correct Answer: B

https://www.slideshare.net/AmazonWebServices/train-models-on-amazon-sagemaker-using-data-not-from-amazon-s3-aim419-aws-reinvent-2018

QUESTION 2

An analytics company has an Amazon SageMaker hosted endpoint for an image classification model. The model is a custom-built convolutional neural network (CNN) and uses the PyTorch deep learning framework. The company wants to increase throughput and decrease latency for customers that use the model.

Which solution will meet these requirements MOST cost-effectively?

A. Use Amazon Elastic Inference on the SageMaker hosted endpoint.

B. Retrain the CNN with more layers and a larger dataset.

C. Retrain the CNN with more layers and a smaller dataset.

D. Choose a SageMaker instance type that has multiple GPUs.

Correct Answer: A

QUESTION 3

A company is converting a large number of unstructured paper receipts into images. The company wants to create a model based on natural language processing (NLP) to find relevant entities such as date, location, and notes, as well as some custom entities such as receipt numbers.

The company is using optical character recognition (OCR) to extract text for data labeling. However, documents are in different structures and formats, and the company is facing challenges with setting up the manual workflows for each document type. Additionally, the company trained a named entity recognition (NER) model for custom entity detection using a small sample size. This model has a very low confidence score and will require retraining with a large dataset.

Which solution for text extraction and entity detection will require the LEAST amount of effort?

A. Extract text from receipt images by using Amazon Textract. Use the Amazon SageMaker BlazingText algorithm to train on the text for entities and custom entities.

B. Extract text from receipt images by using a deep learning OCR model from the AWS Marketplace. Use the NER deep learning model to extract entities.

C. Extract text from receipt images by using Amazon Textract. Use Amazon Comprehend for entity detection, and use Amazon Comprehend custom entity recognition for custom entity detection.

D. Extract text from receipt images by using a deep learning OCR model from the AWS Marketplace. Use Amazon Comprehend for entity detection, and use Amazon Comprehend custom entity recognition for custom entity detection.

Correct Answer: C

Reference: https://aws.amazon.com/blogs/machine-learning/building-an-nlp-powered-search-index-with-amazon-textract-and-amazon-comprehend/

QUESTION 4

A company wants to forecast the daily price of newly launched products based on 3 years of data for older product prices, sales, and rebates. The time-series data has irregular timestamps and is missing some values.

Data scientist must build a dataset to replace the missing values. The data scientist needs a solution that resamptes the data daily and exports the data for further modeling.

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

A. Use Amazon EMR Serveriess with PySpark.

B. Use AWS Glue DataBrew.

C. Use Amazon SageMaker Studio Data Wrangler.

D. Use Amazon SageMaker Studio Notebook with Pandas.

Correct Answer: C

Amazon SageMaker Studio Data Wrangler is a visual data preparation tool that enables users to clean and normalize data without writing any code. Using Data Wrangler, the data scientist can easily import the time-series data from various sources, such as Amazon S3, Amazon Athena, or Amazon Redshift. Data Wrangler can automatically generate data insights and quality reports, which can help identify and fix missing values, outliers, and anomalies in the data. Data Wrangler also provides over 250 built-in transformations, such as resampling, interpolation, aggregation, and filtering, which can be applied to the data with a point-and-click interface. Data Wrangler can also export the prepared data to different destinations, such as Amazon S3, Amazon SageMaker Feature Store, or Amazon SageMaker Pipelines, for further modeling and analysis. Data Wrangler is integrated with Amazon SageMaker Studio, a web- based IDE for machine learning, which makes it easy to access and use the tool. Data Wrangler is a serverless and fully managed service, which means the data scientist does not need to provision, configure, or manage any infrastructure or clusters. Option A is incorrect because Amazon EMR Serverless is a serverless option for running big data analytics applications using open-source frameworks, such as Apache Spark. However, using Amazon EMR Serverless would require the data scientist to write PySpark code to perform the data preparation tasks, such as resampling, imputation, and aggregation. This would require more implementation effort than using Data Wrangler, which provides a visual and code-free interface for data preparation. Option B is incorrect because AWS Glue DataBrew is another visual data preparation tool that can be used to clean and normalize data without writing code. However, DataBrew does not support time-series data as a data type, and does not provide built-in transformations for resampling, interpolation, or aggregation of time-series data. Therefore, using DataBrew would not meet the requirements of the use case. Option D

is incorrect because using Amazon SageMaker Studio Notebook with Pandas would also require the data scientist to write Python code to perform the data preparation tasks. Pandas is a popular Python library for data analysis and manipulation, which supports time-series data and provides various methods for resampling, interpolation, and aggregation. However, using Pandas would require more implementation effort than using Data Wrangler, which provides a visual and code-free interface for data preparation. References:

1: Amazon SageMaker Data Wrangler documentation

2: Amazon EMR Serverless documentation

3: AWS Glue DataBrew documentation

4: Pandas documentation

QUESTION 5

A data scientist is working on a forecast problem by using a dataset that consists of .csv files that are stored in Amazon S3. The files contain a timestamp variable in the following format: March 1st, 2020, 08:14pm There is a hypothesis about seasonal differences in the dependent variable. This number could be higher or lower for weekdays because some days and hours present varying values, so the day of the week, month, or hour could be an

important factor. As a result, the data scientist needs to transform the timestamp into weekdays, month, and day as three separate variables to conduct an analysis. Which solution requires the LEAST operational overhead to create a new dataset with the added features?

A. Create an Amazon EMR cluster. Develop PySpark code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.

B. Create a processing job in Amazon SageMaker. Develop Python code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.

C. Create a new flow in Amazon SageMaker Data Wrangler. Import the S3 file, use the Featurize date/time transform to generate the new variables, and save the dataset as a new file in Amazon S3.

D. Create an AWS Glue job. Develop code that can read the timestamp variable as a string, transform and create the new variables, and save the dataset as a new file in Amazon S3.

Correct Answer: C