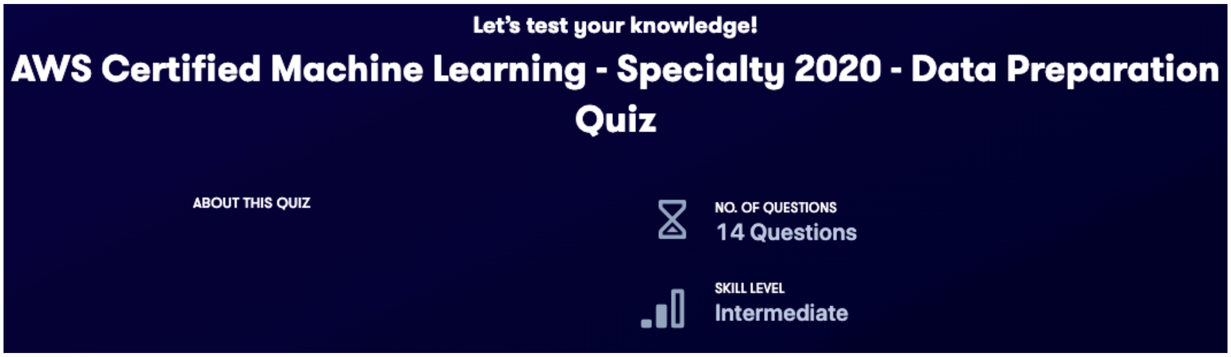
Cloud Guru - 3 - Data Preparation Quiz

https://acloud.guru/course/aws-certified-machine-learning-specialty/learn/f399041c-e040-9d9a-d60c-35317e0e9a97/chapter-4/watch?backUrl=~2Fcourses



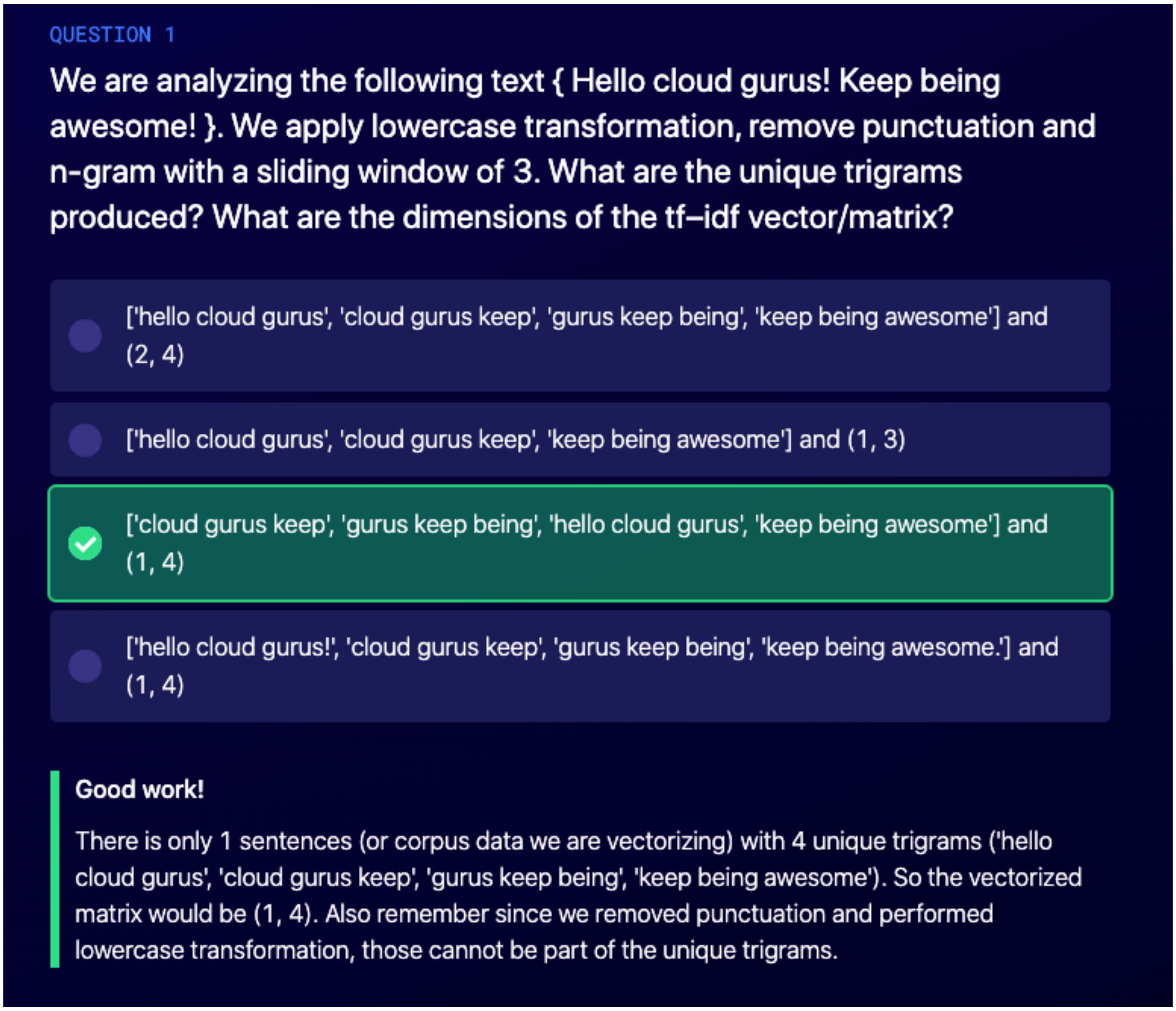
Let's test your knowledge!

AWS Certified Machine Learning - Specialty 2020 - Data Preparation Quiz

ABOUT THIS QUIZ

NO. OF QUESTIONS
14 Questions

SKILL LEVEL
Intermediate



QUESTION 1

We are analyzing the following text { Hello cloud gurus! Keep being awesome! }. We apply lowercase transformation, remove punctuation and n-gram with a sliding window of 3. What are the unique trigrams produced? What are the dimensions of the tf–idf vector/matrix?

○ ['hello cloud gurus', 'cloud gurus keep', 'gurus keep being', 'keep being awesome'] and (2, 4)

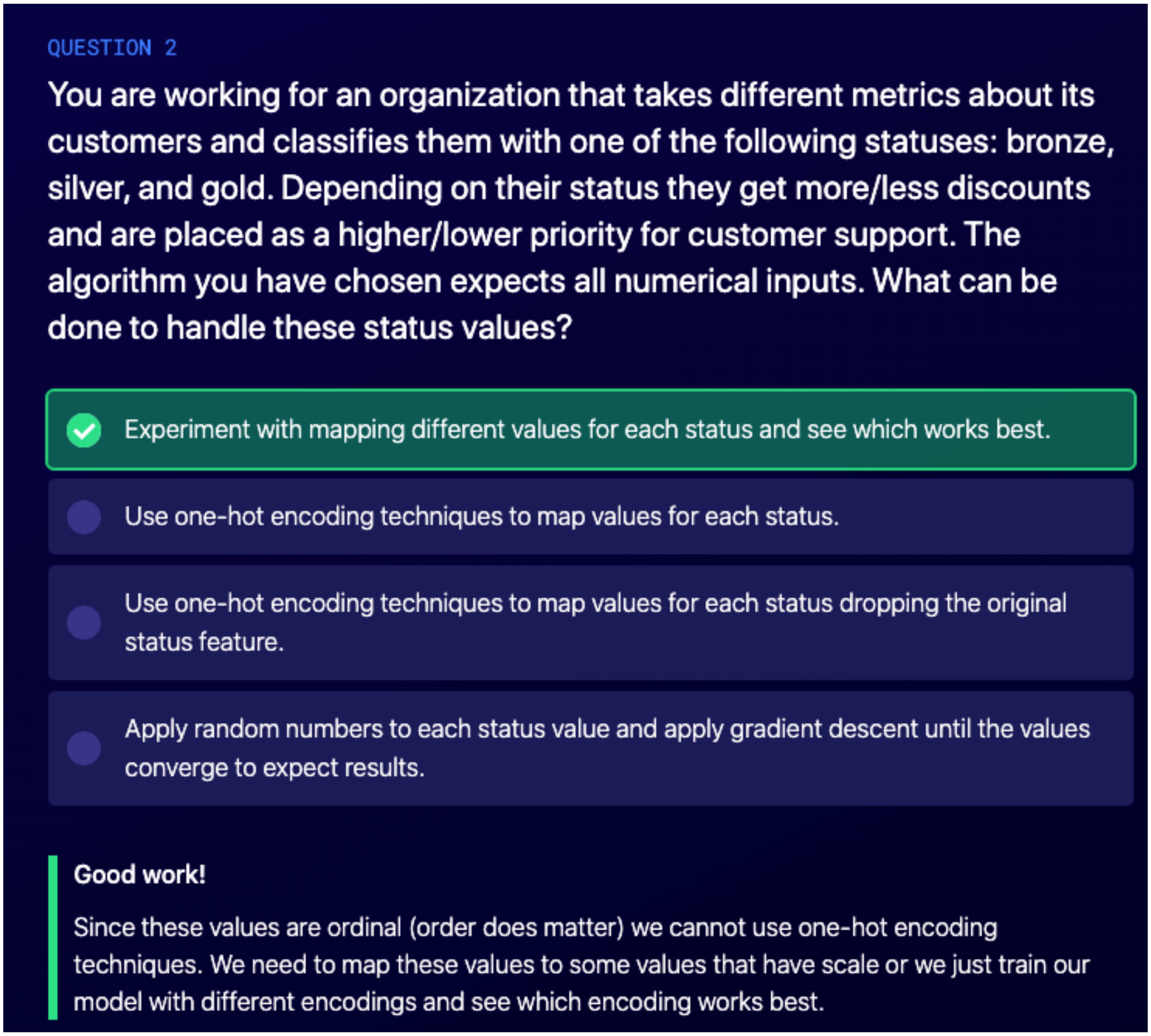
○ ['hello cloud gurus', 'cloud gurus keep', 'keep being awesome'] and (1, 3)

✓ ['cloud gurus keep', 'gurus keep being', 'hello cloud gurus', 'keep being awesome'] and (1, 4)

○ ['hello cloud gurus!', 'cloud gurus keep', 'gurus keep being', 'keep being awesome.'] and (1, 4)

Good work!

There is only 1 sentences (or corpus data we are vectorizing) with 4 unique trigrams ('hello cloud gurus', 'cloud gurus keep', 'gurus keep being', 'keep being awesome'). So the vectorized matrix would be (1, 4). Also remember since we removed punctuation and performed lowercase transformation, those cannot be part of the unique trigrams.



QUESTION 2

You are working for an organization that takes different metrics about its customers and classifies them with one of the following statuses: bronze, silver, and gold. Depending on their status they get more/less discounts and are placed as a higher/lower priority for customer support. The algorithm you have chosen expects all numerical inputs. What can be done to handle these status values?

✓ Experiment with mapping different values for each status and see which works best.

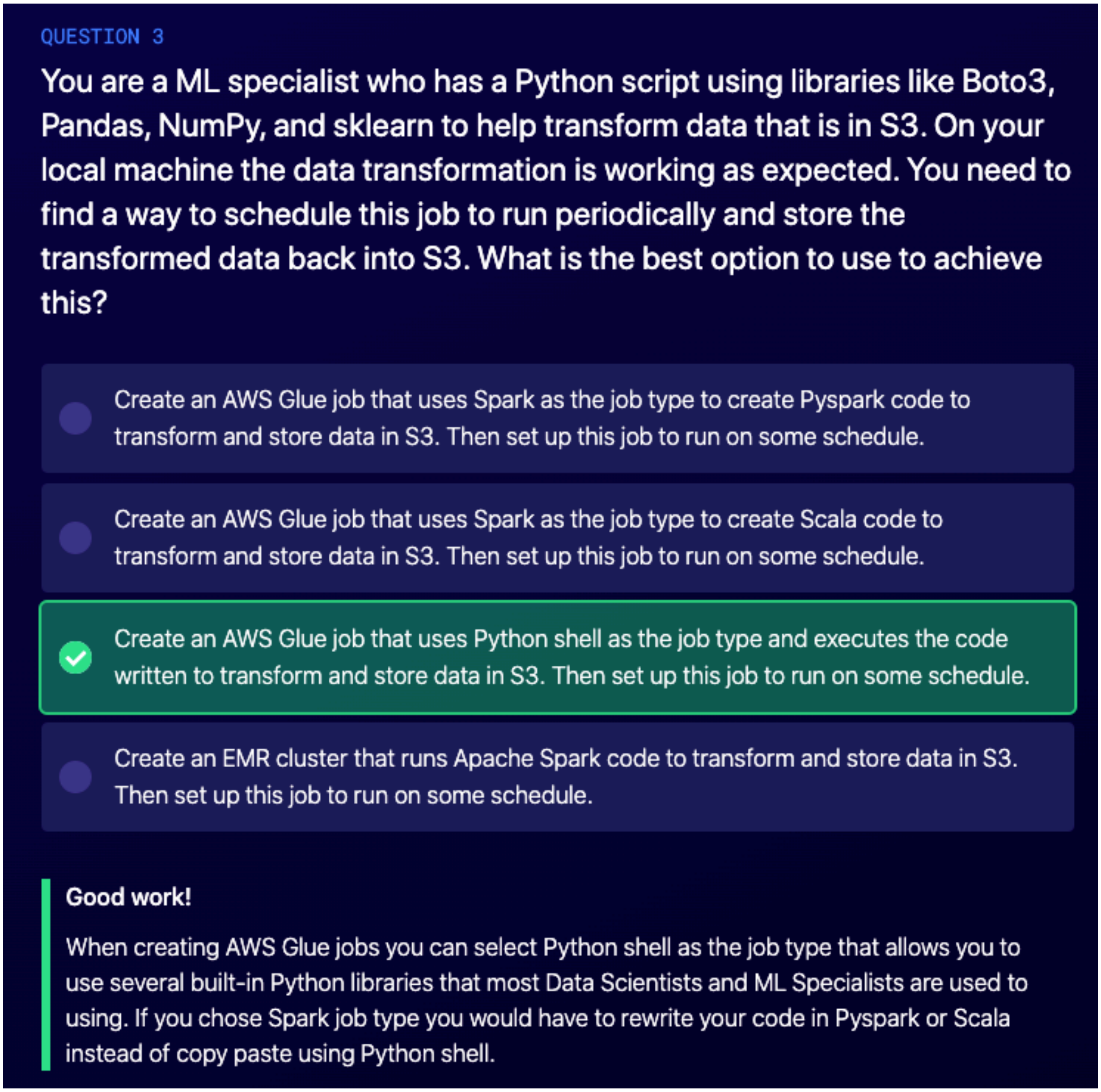
○ Use one-hot encoding techniques to map values for each status.

○ Use one-hot encoding techniques to map values for each status dropping the original status feature.

○ Apply random numbers to each status value and apply gradient descent until the values converge to expect results.

Good work!

Since these values are ordinal (order does matter) we cannot use one-hot encoding techniques. We need to map these values to some values that have scale or we just train our model with different encodings and see which encoding works best.



QUESTION 3

You are a ML specialist who has a Python script using libraries like Boto3, Pandas, NumPy, and sklearn to help transform data that is in S3. On your local machine the data transformation is working as expected. You need to find a way to schedule this job to run periodically and store the transformed data back into S3. What is the best option to use to achieve this?

○ Create an AWS Glue job that uses Spark as the job type to create Pyspark code to transform and store data in S3. Then set up this job to run on some schedule.

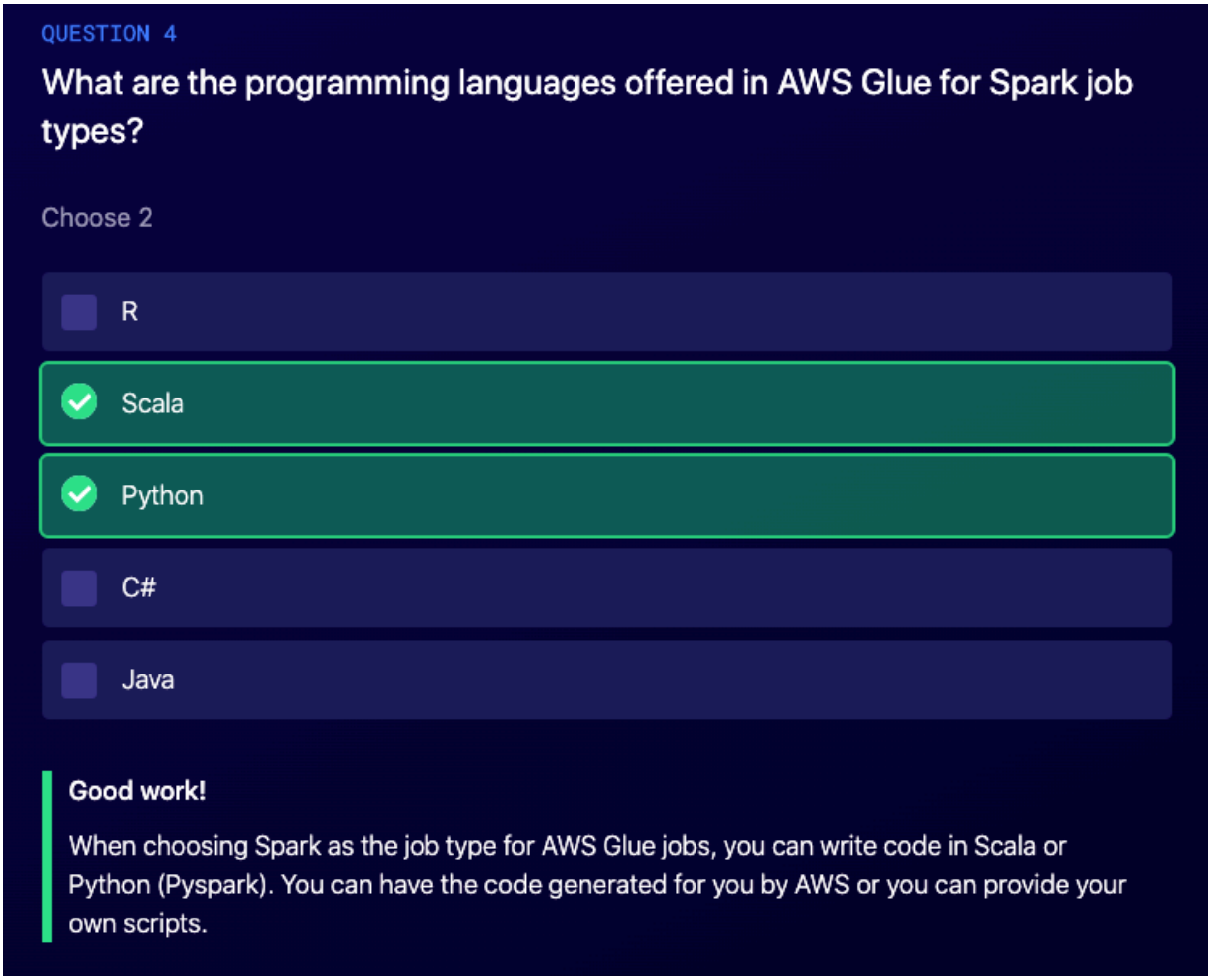
○ Create an AWS Glue job that uses Spark as the job type to create Scala code to transform and store data in S3. Then set up this job to run on some schedule.

✓ Create an AWS Glue job that uses Python shell as the job type and executes the code written to transform and store data in S3. Then set up this job to run on some schedule.

○ Create an EMR cluster that runs Apache Spark code to transform and store data in S3. Then set up this job to run on some schedule.

Good work!

When creating AWS Glue jobs you can select Python shell as the job type that allows you to use several built-in Python libraries that most Data Scientists and ML Specialists are used to using. If you chose Spark job type you would have to rewrite your code in Pyspark or Scala instead of copy paste using Python shell.



QUESTION 4

What are the programming languages offered in AWS Glue for Spark job types?

Choose 2

□ R

✓ Scala

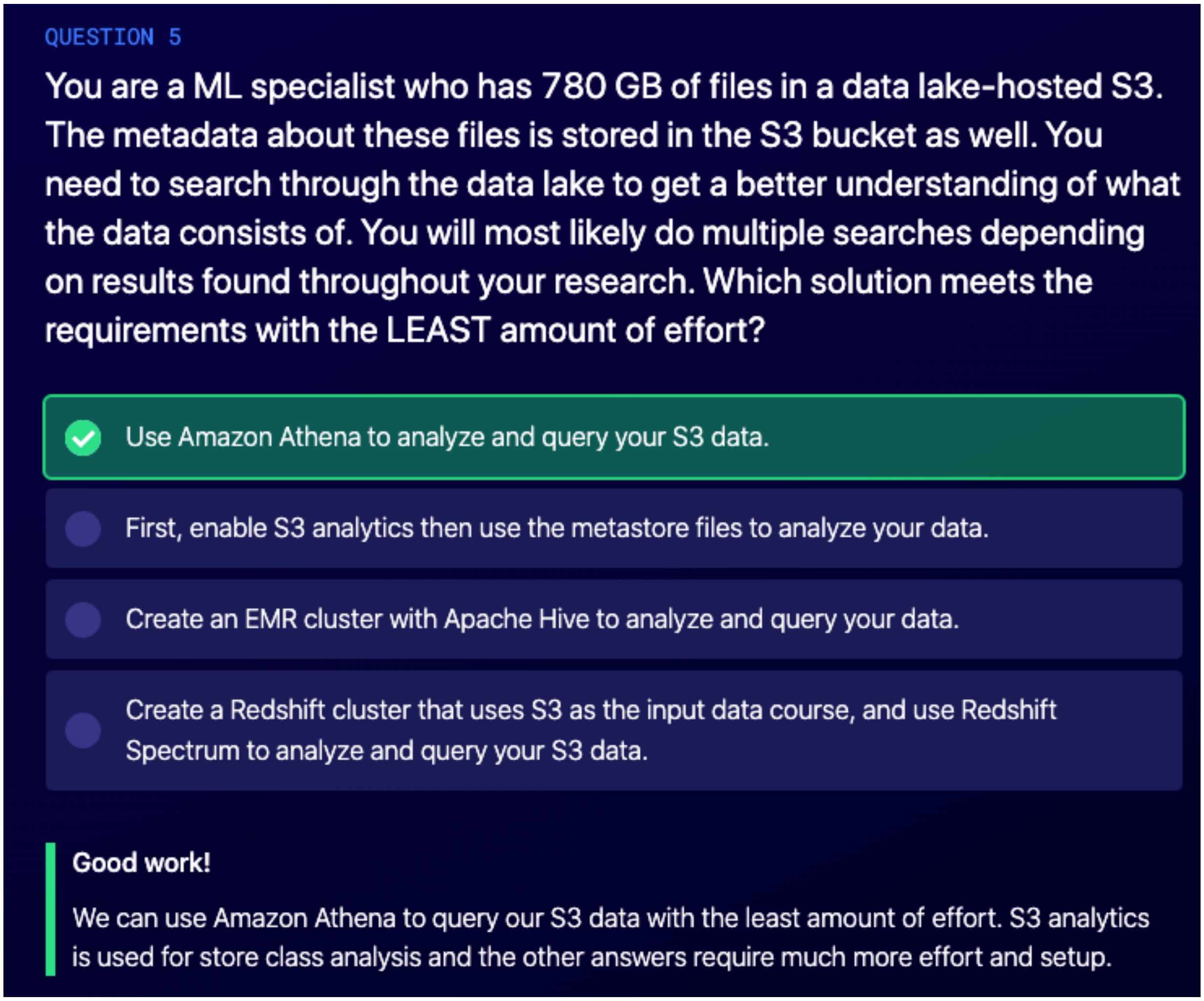
✓ Python

□ C#

□ Java

Good work!

When choosing Spark as the job type for AWS Glue jobs, you can write code in Scala or Python (Pyspark). You can have the code generated for you by AWS or you can provide your own scripts.



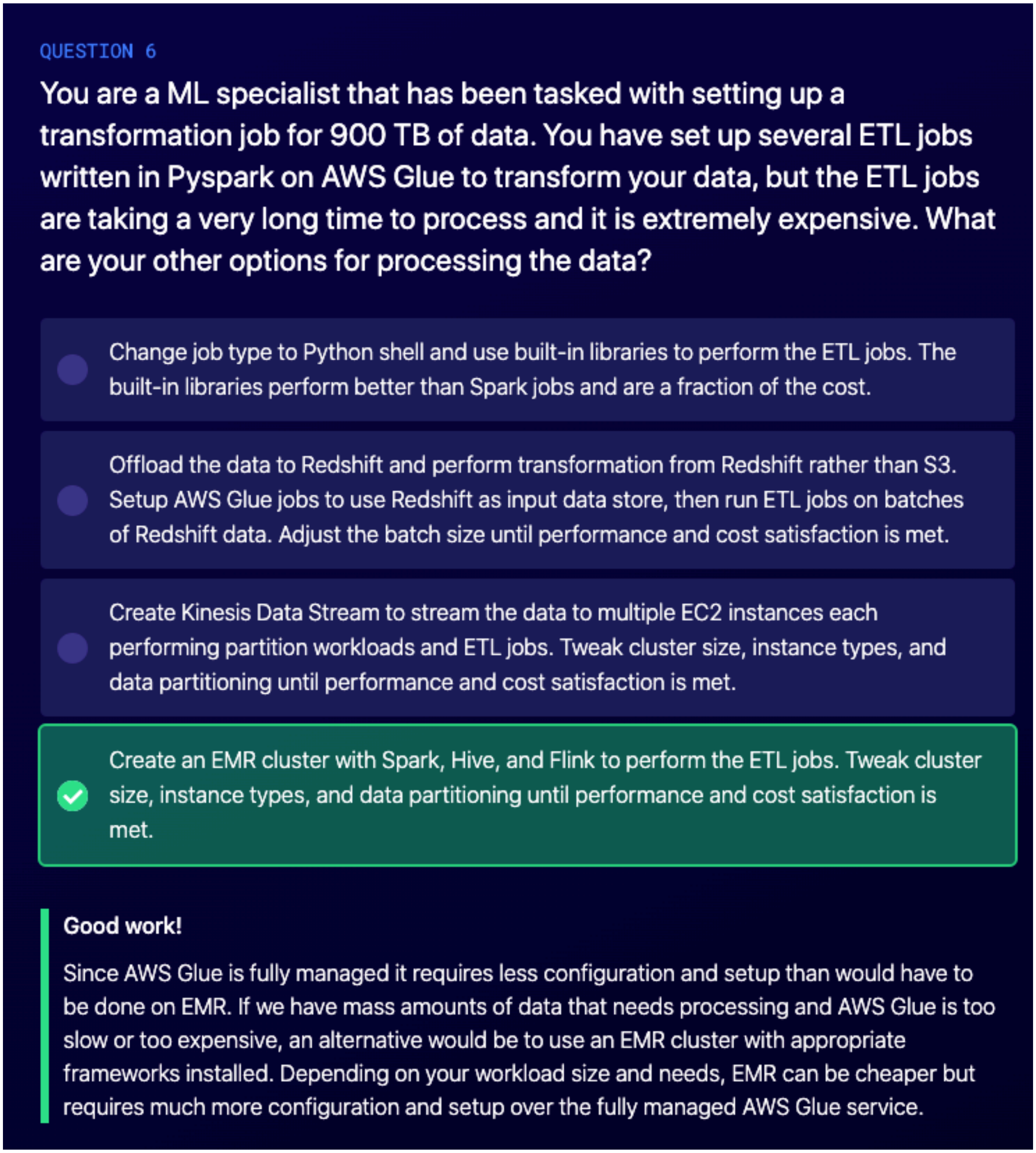
QUESTION 5

You are a ML specialist who has 780 GB of files in a data lake-hosted S3. The metadata about these files is stored in the S3 bucket as well. You need to search through the data lake to get a better understanding of what the data consists of. You will most likely do multiple searches depending on results found throughout your research. Which solution meets the requirements with the LEAST amount of effort?

- ✓ Use Amazon Athena to analyze and query your S3 data.
- ○ First, enable S3 analytics then use the metastore files to analyze your data.
- ○ Create an EMR cluster with Apache Hive to analyze and query your data.
- ○ Create a Redshift cluster that uses S3 as the input data course, and use Redshift Spectrum to analyze and query your S3 data.

Good work!

We can use Amazon Athena to query our S3 data with the least amount of effort. S3 analytics is used for store class analysis and the other answers require much more effort and setup.



QUESTION 6

You are a ML specialist that has been tasked with setting up a transformation job for 900 TB of data. You have set up several ETL jobs written in Pyspark on AWS Glue to transform your data, but the ETL jobs are taking a very long time to process and it is extremely expensive. What are your other options for processing the data?

○ Change job type to Python shell and use built-in libraries to perform the ETL jobs. The built-in libraries perform better than Spark jobs and are a fraction of the cost.

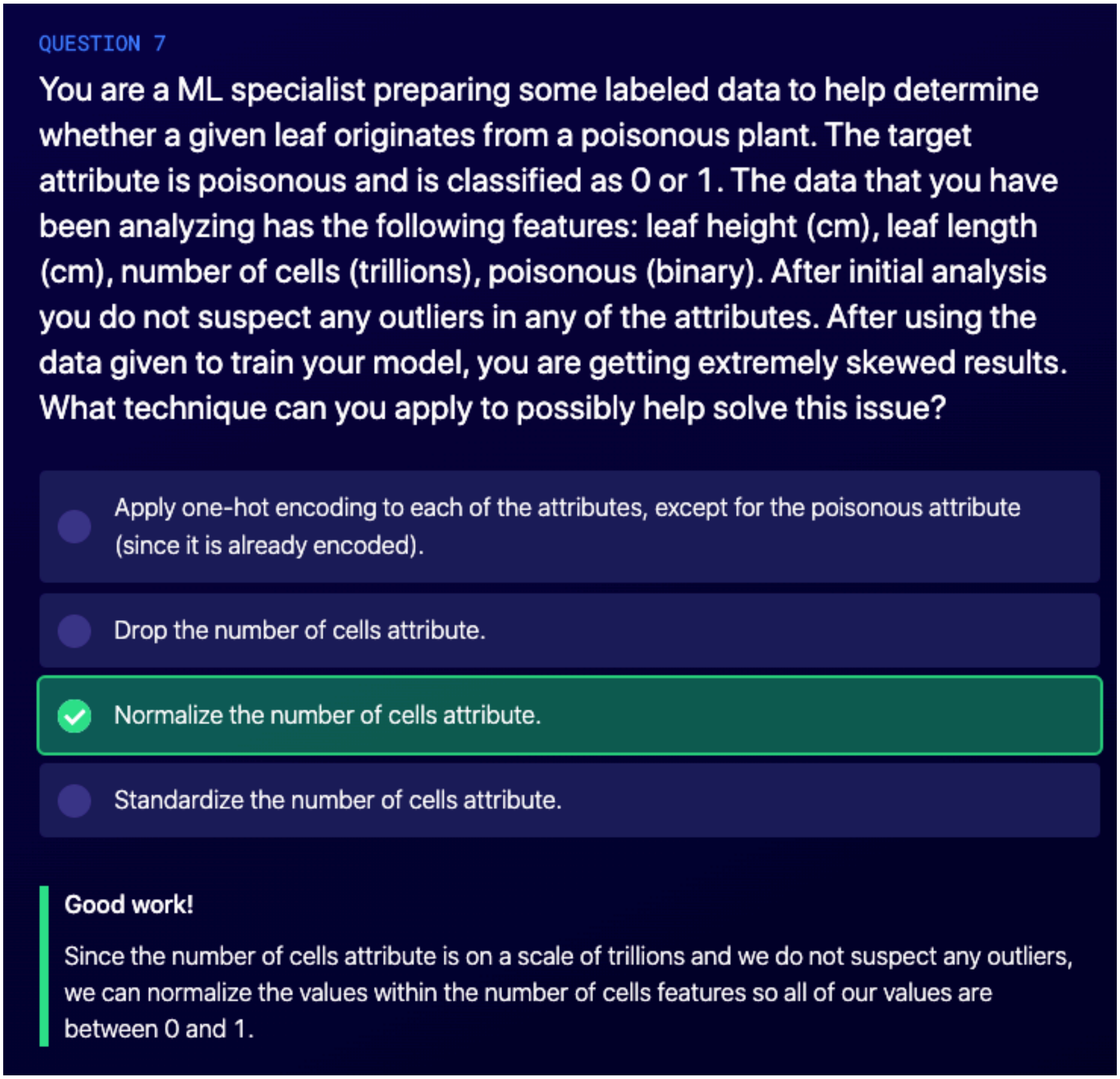
○ Offload the data to Redshift and perform transformation from Redshift rather than S3. Setup AWS Glue jobs to use Redshift as input data store, then run ETL jobs on batches of Redshift data. Adjust the batch size until performance and cost satisfaction is met.

○ Create Kinesis Data Stream to stream the data to multiple EC2 instances each performing partition workloads and ETL jobs. Tweak cluster size, instance types, and data partitioning until performance and cost satisfaction is met.

✓ Create an EMR cluster with Spark, Hive, and Flink to perform the ETL jobs. Tweak cluster size, instance types, and data partitioning until performance and cost satisfaction is met.

Good work!

Since AWS Glue is fully managed it requires less configuration and setup than would have to be done on EMR. If we have mass amounts of data that needs processing and AWS Glue is too slow or too expensive, an alternative would be to use an EMR cluster with appropriate frameworks installed. Depending on your workload size and needs, EMR can be cheaper but requires much more configuration and setup over the fully managed AWS Glue service.



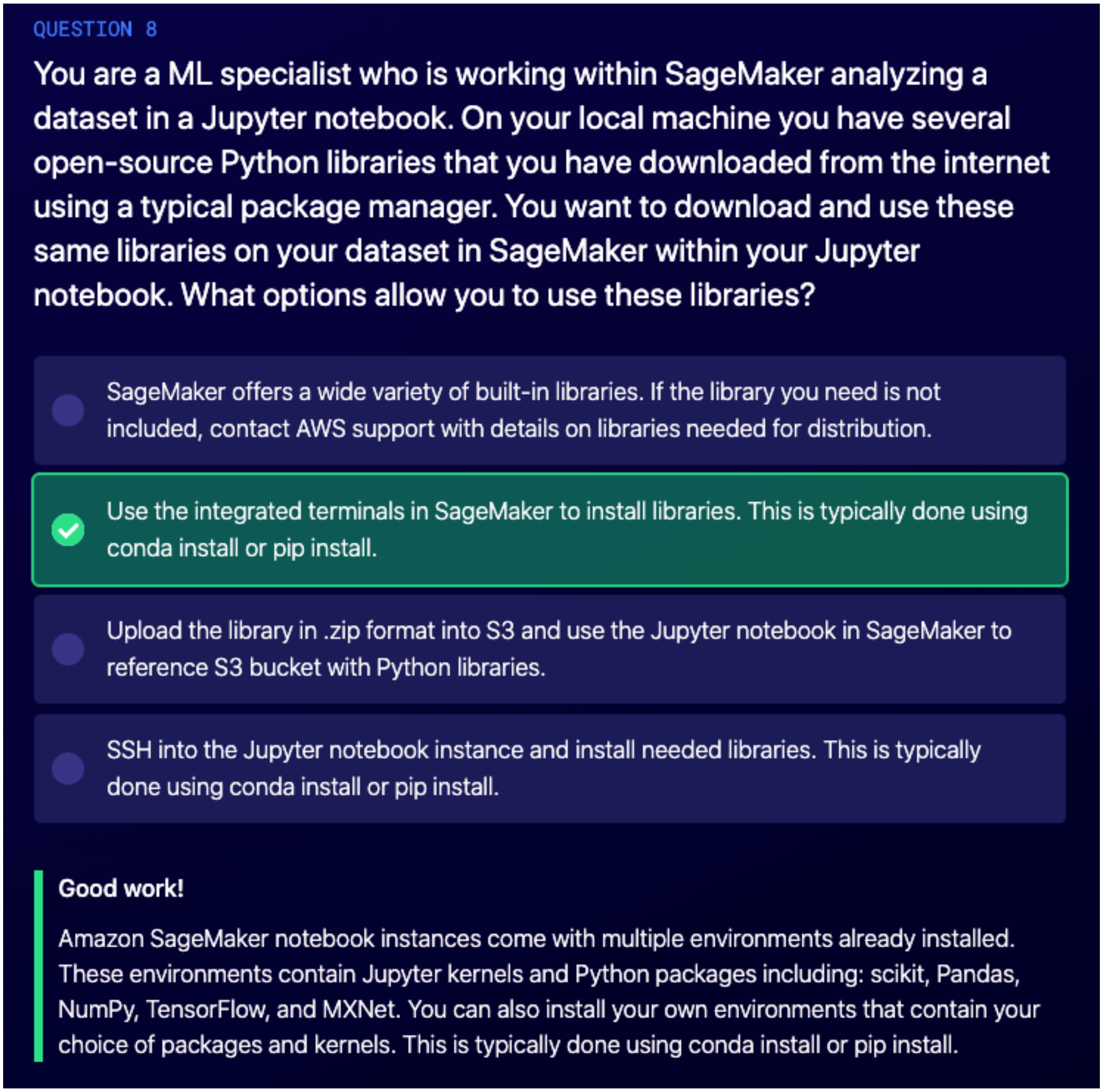
QUESTION 7

You are a ML specialist preparing some labeled data to help determine whether a given leaf originates from a poisonous plant. The target attribute is poisonous and is classified as 0 or 1. The data that you have been analyzing has the following features: leaf height (cm), leaf length (cm), number of cells (trillions), poisonous (binary). After initial analysis you do not suspect any outliers in any of the attributes. After using the data given to train your model, you are getting extremely skewed results. What technique can you apply to possibly help solve this issue?

- ○ Apply one-hot encoding to each of the attributes, except for the poisonous attribute (since it is already encoded).
- ○ Drop the number of cells attribute.
- ✓ Normalize the number of cells attribute.
- ○ Standardize the number of cells attribute.

Good work!

Since the number of cells attribute is on a scale of trillions and we do not suspect any outliers, we can normalize the values within the number of cells features so all of our values are between 0 and 1.



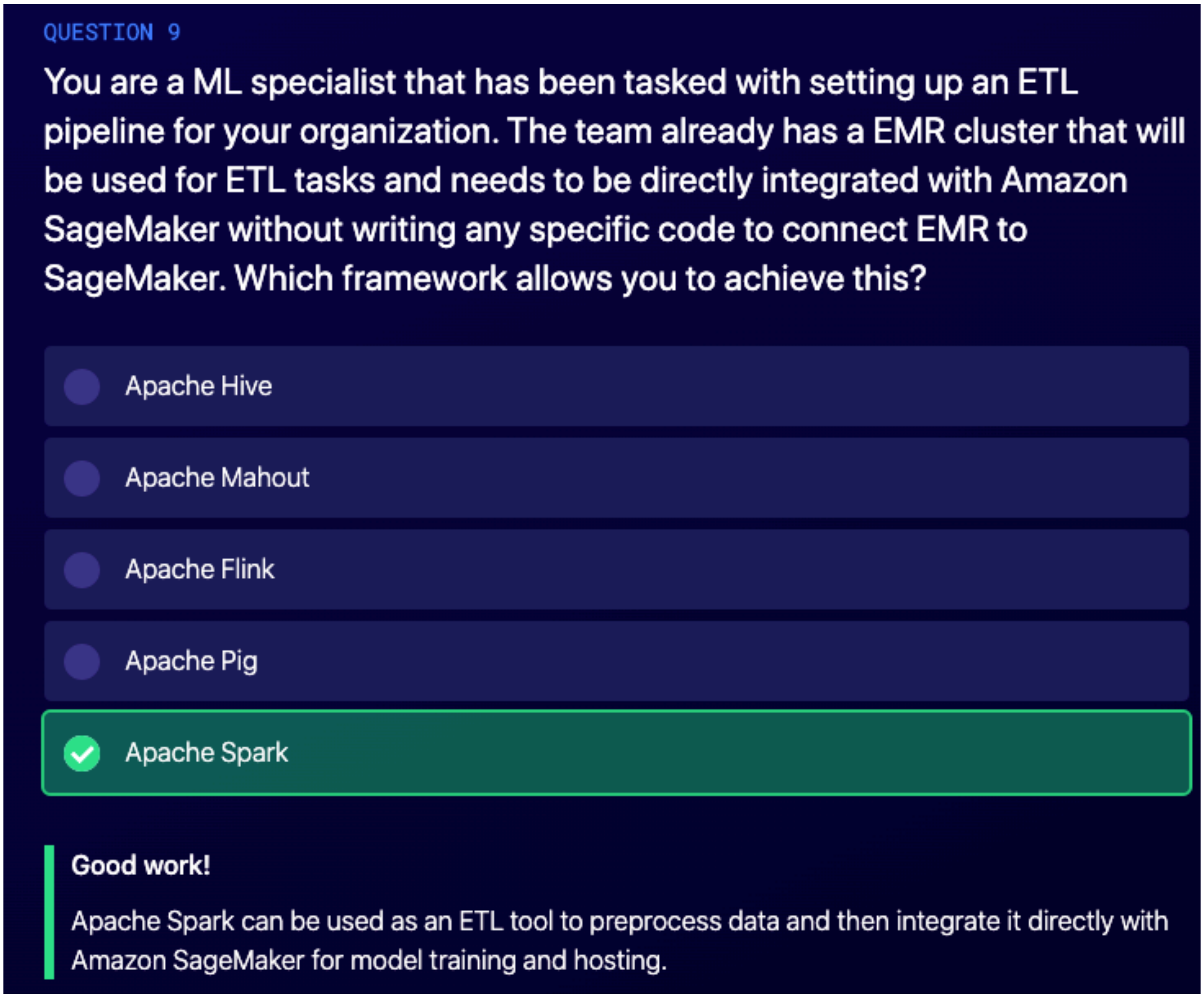
QUESTION 8

You are a ML specialist who is working within SageMaker analyzing a dataset in a Jupyter notebook. On your local machine you have several open-source Python libraries that you have downloaded from the internet using a typical package manager. You want to download and use these same libraries on your dataset in SageMaker within your Jupyter notebook. What options allow you to use these libraries?

- ○ SageMaker offers a wide variety of built-in libraries. If the library you need is not included, contact AWS support with details on libraries needed for distribution.
- ✓ Use the integrated terminals in SageMaker to install libraries. This is typically done using conda install or pip install.
- ○ Upload the library in .zip format into S3 and use the Jupyter notebook in SageMaker to reference S3 bucket with Python libraries.
- ○ SSH into the Jupyter notebook instance and install needed libraries. This is typically done using conda install or pip install.

Good work!

Amazon SageMaker notebook instances come with multiple environments already installed. These environments contain Jupyter kernels and Python packages including: scikit, Pandas, NumPy, TensorFlow, and MXNet. You can also install your own environments that contain your choice of packages and kernels. This is typically done using conda install or pip install.



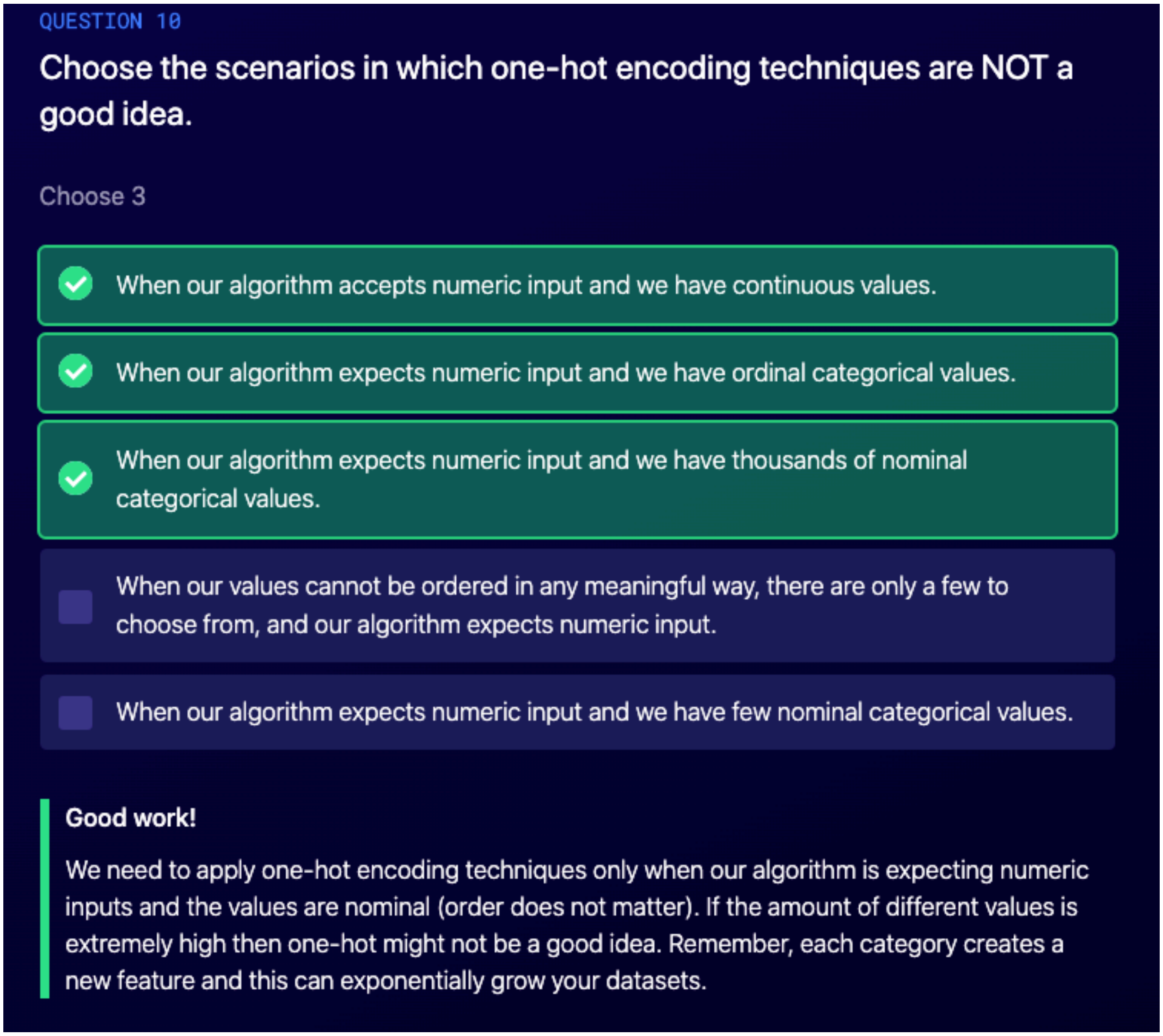
QUESTION 9

You are a ML specialist that has been tasked with setting up an ETL pipeline for your organization. The team already has a EMR cluster that will be used for ETL tasks and needs to be directly integrated with Amazon SageMaker without writing any specific code to connect EMR to SageMaker. Which framework allows you to achieve this?

- ○ Apache Hive
- ○ Apache Mahout
- ○ Apache Flink
- ○ Apache Pig
- ✓ Apache Spark

Good work!

Apache Spark can be used as an ETL tool to preprocess data and then integrate it directly with Amazon SageMaker for model training and hosting.



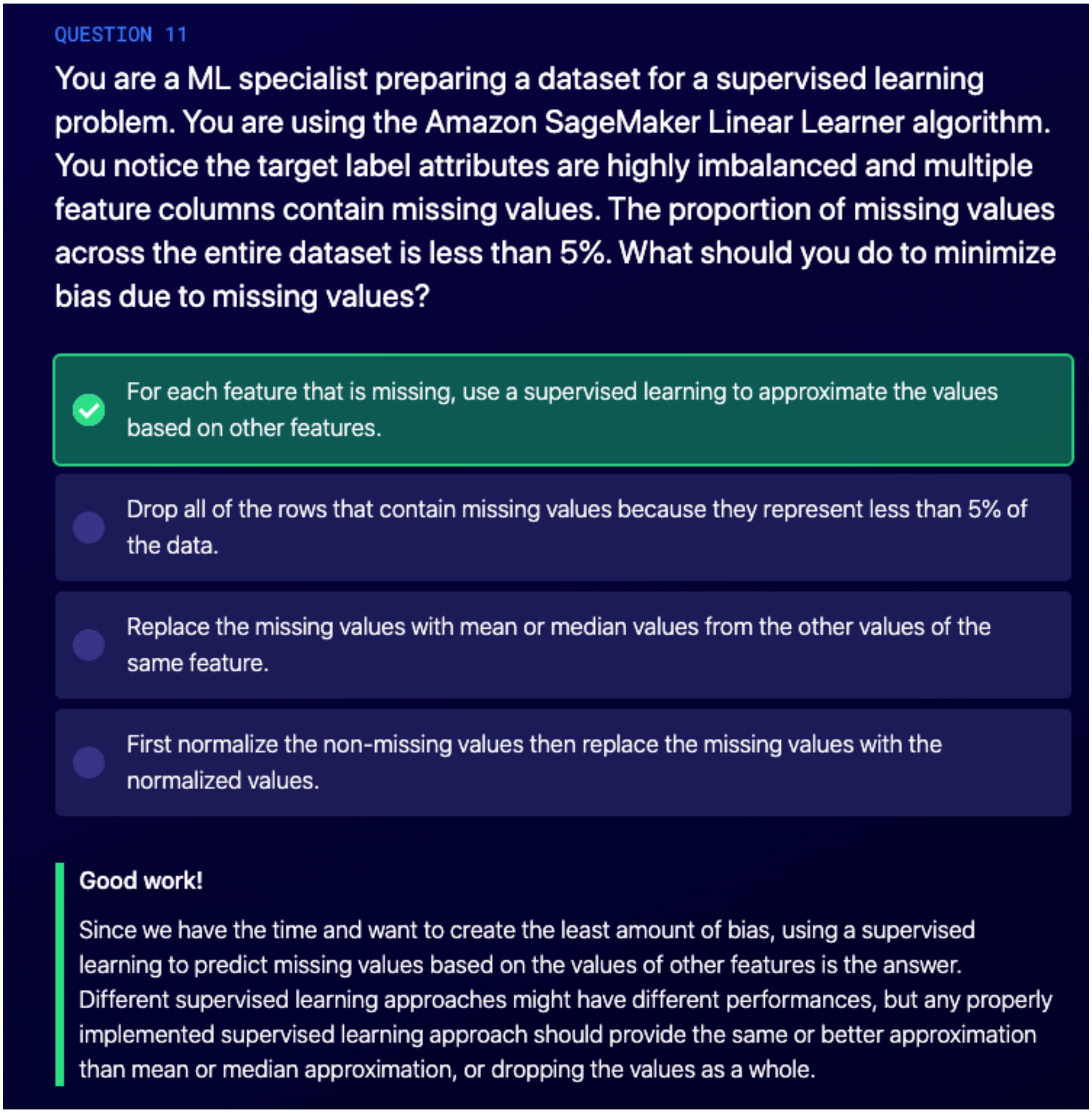
QUESTION 10

Choose the scenarios in which one-hot encoding techniques are NOT a good idea.

Choose 3

- ☑ When our algorithm accepts numeric input and we have continuous values.
- ☑ When our algorithm expects numeric input and we have ordinal categorical values.
- ☑ When our algorithm expects numeric input and we have thousands of nominal categorical values.
- ☐ When our values cannot be ordered in any meaningful way, there are only a few to choose from, and our algorithm expects numeric input.
- ☐ When our algorithm expects numeric input and we have few nominal categorical values.

> **Good work!**
>
> We need to apply one-hot encoding techniques only when our algorithm is expecting numeric inputs and the values are nominal (order does not matter). If the amount of different values is extremely high then one-hot might not be a good idea. Remember, each category creates a new feature and this can exponentially grow your datasets.



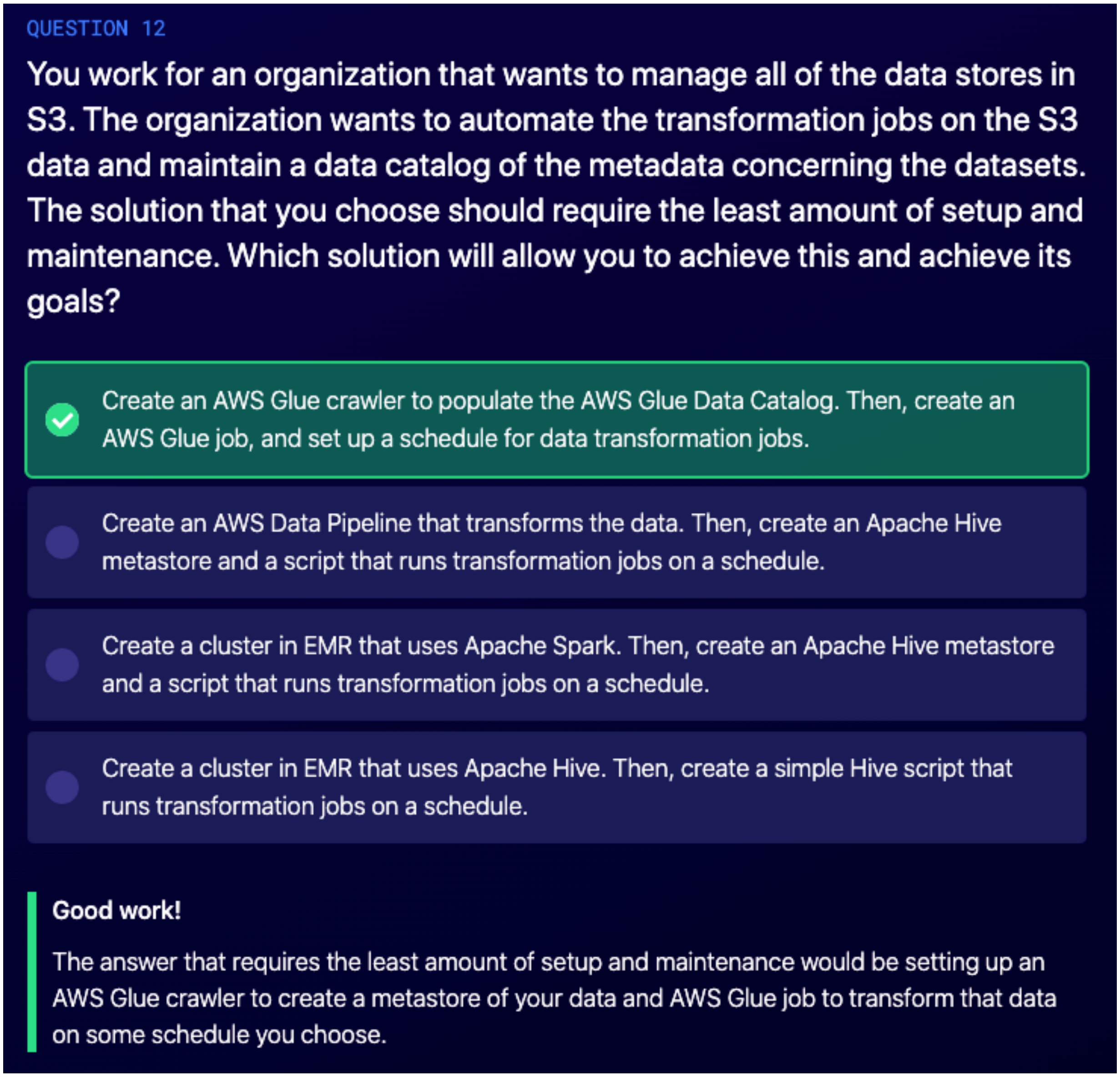
QUESTION 11

You are a ML specialist preparing a dataset for a supervised learning problem. You are using the Amazon SageMaker Linear Learner algorithm. You notice the target label attributes are highly imbalanced and multiple feature columns contain missing values. The proportion of missing values across the entire dataset is less than 5%. What should you do to minimize bias due to missing values?

- ✓ For each feature that is missing, use a supervised learning to approximate the values based on other features.
- ○ Drop all of the rows that contain missing values because they represent less than 5% of the data.
- ○ Replace the missing values with mean or median values from the other values of the same feature.
- ○ First normalize the non-missing values then replace the missing values with the normalized values.

Good work!

Since we have the time and want to create the least amount of bias, using a supervised learning to predict missing values based on the values of other features is the answer. Different supervised learning approaches might have different performances, but any properly implemented supervised learning approach should provide the same or better approximation than mean or median approximation, or dropping the values as a whole.



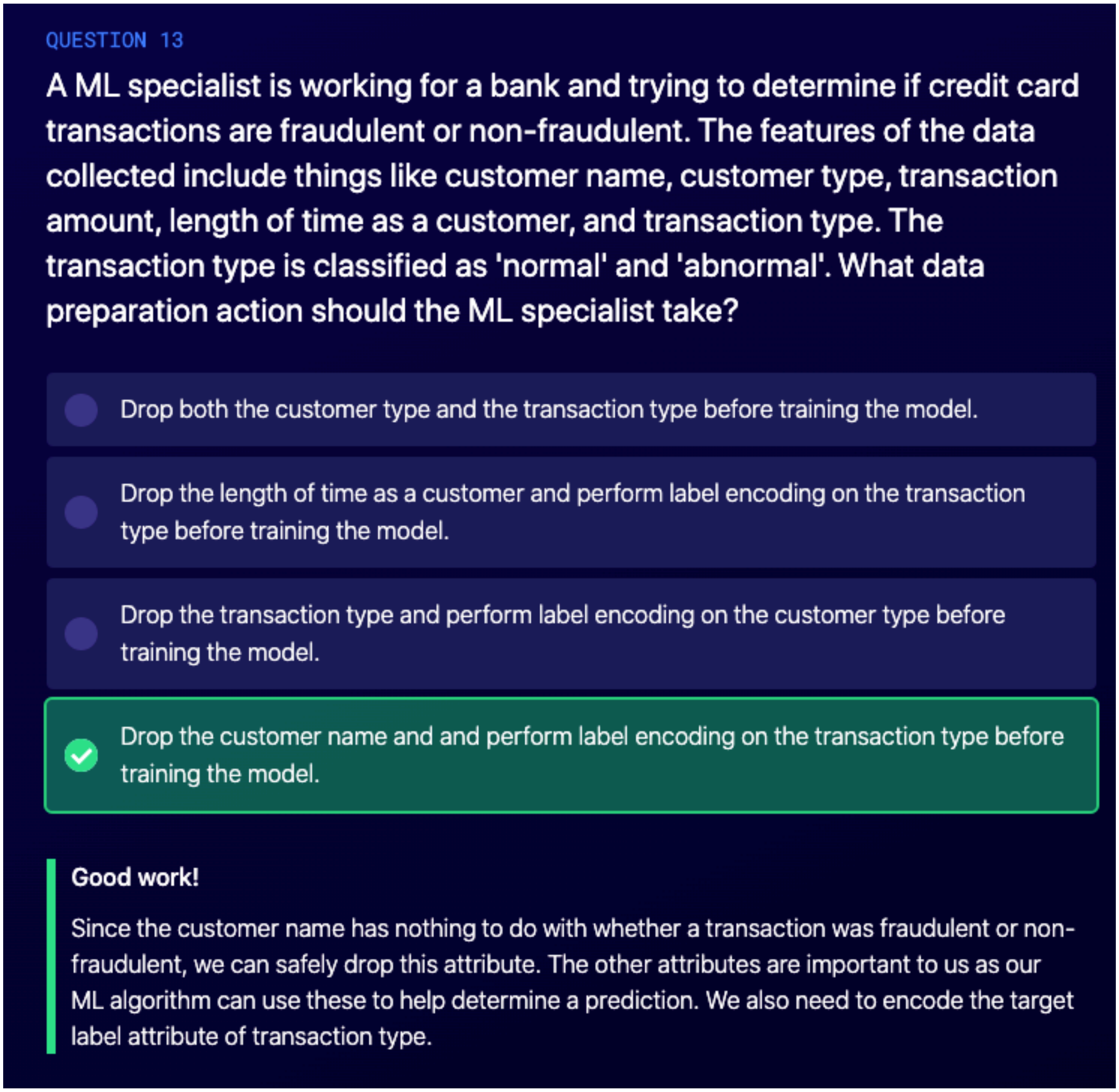
QUESTION 12

You work for an organization that wants to manage all of the data stores in S3. The organization wants to automate the transformation jobs on the S3 data and maintain a data catalog of the metadata concerning the datasets. The solution that you choose should require the least amount of setup and maintenance. Which solution will allow you to achieve this and achieve its goals?

- ✓ Create an AWS Glue crawler to populate the AWS Glue Data Catalog. Then, create an AWS Glue job, and set up a schedule for data transformation jobs.
- ○ Create an AWS Data Pipeline that transforms the data. Then, create an Apache Hive metastore and a script that runs transformation jobs on a schedule.
- ○ Create a cluster in EMR that uses Apache Spark. Then, create an Apache Hive metastore and a script that runs transformation jobs on a schedule.
- ○ Create a cluster in EMR that uses Apache Hive. Then, create a simple Hive script that runs transformation jobs on a schedule.

Good work!

The answer that requires the least amount of setup and maintenance would be setting up an AWS Glue crawler to create a metastore of your data and AWS Glue job to transform that data on some schedule you choose.



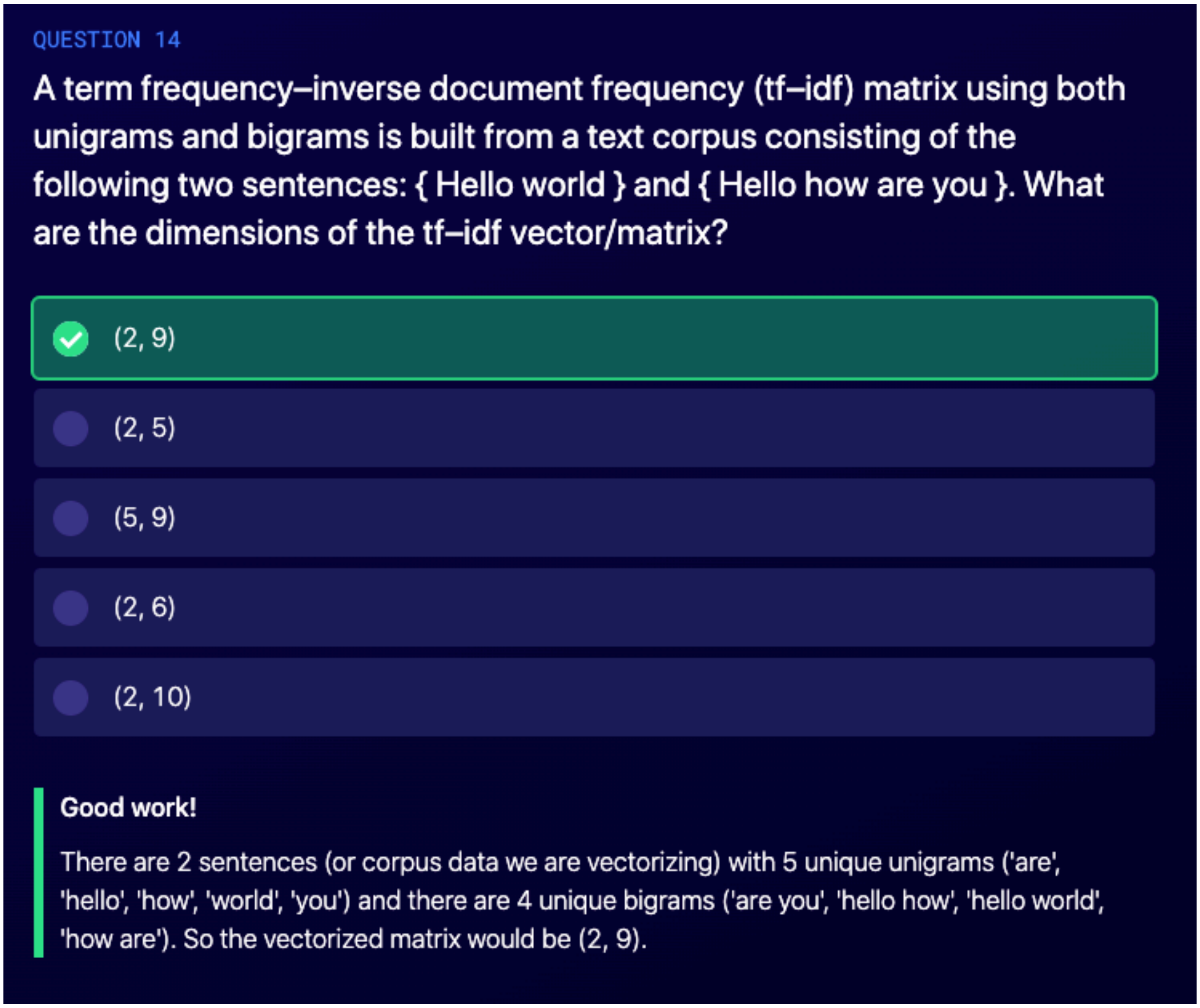
QUESTION 13

A ML specialist is working for a bank and trying to determine if credit card transactions are fraudulent or non-fraudulent. The features of the data collected include things like customer name, customer type, transaction amount, length of time as a customer, and transaction type. The transaction type is classified as 'normal' and 'abnormal'. What data preparation action should the ML specialist take?

- ○ Drop both the customer type and the transaction type before training the model.
- ○ Drop the length of time as a customer and perform label encoding on the transaction type before training the model.
- ○ Drop the transaction type and perform label encoding on the customer type before training the model.
- ✓ Drop the customer name and and perform label encoding on the transaction type before training the model.

Good work!

Since the customer name has nothing to do with whether a transaction was fraudulent or non-fraudulent, we can safely drop this attribute. The other attributes are important to us as our ML algorithm can use these to help determine a prediction. We also need to encode the target label attribute of transaction type.



QUESTION 14

A term frequency–inverse document frequency (tf–idf) matrix using both unigrams and bigrams is built from a text corpus consisting of the following two sentences: { Hello world } and { Hello how are you }. What are the dimensions of the tf–idf vector/matrix?

- ✓ (2, 9)
- ○ (2, 5)
- ○ (5, 9)
- ○ (2, 6)
- ○ (2, 10)

Good work!

There are 2 sentences (or corpus data we are vectorizing) with 5 unique unigrams ('are', 'hello', 'how', 'world', 'you') and there are 4 unique bigrams ('are you', 'hello how', 'hello world', 'how are'). So the vectorized matrix would be (2, 9).

100%

Congratulations!

You passed AWS Certified Machine Learning - Specialty 2020 - Data Preparation Quiz!