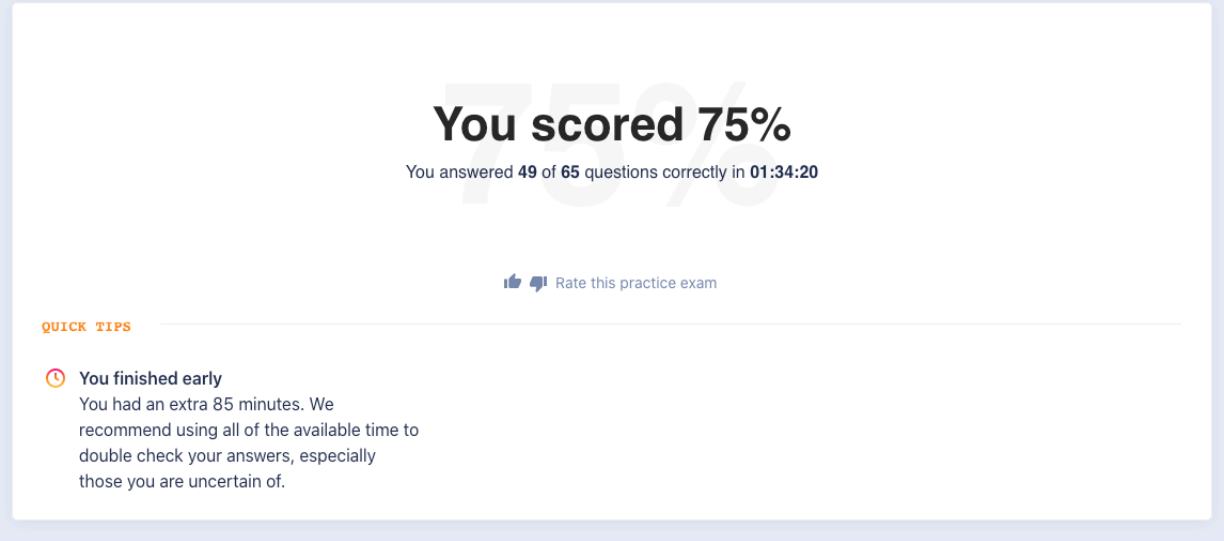
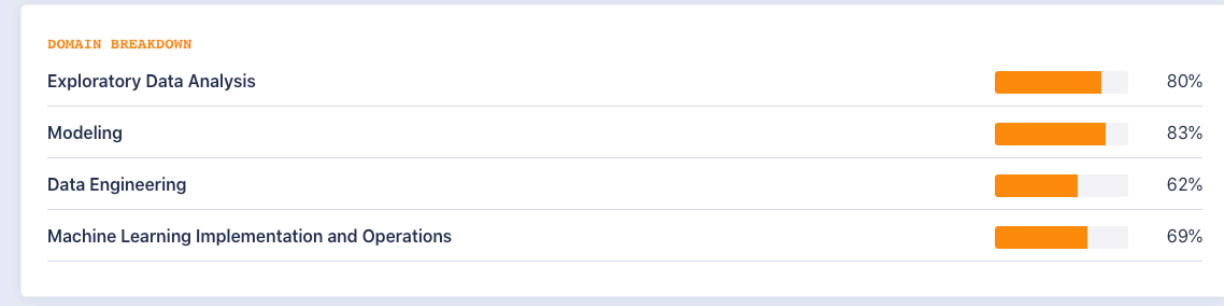
CloudGuru - Practice Exam - Review

https://practice-exam.acloud.guru/aws-certified-machine-learning-specialty/result/b8a3def8-3696-44bc-9a54-34c0ece59992

Reviewing the questions I got wrong





80%

DOMAIN

Exploratory Data Analysis

2.1 Sanitize and prepare data for modeling 2.2 Perform feature engineering 2.3 Analyze and visualize data for machine learning Questions for this domain comprise **24% of the total questions** for this exam.

QUESTIONS

35	53	64	1	4
8	10	17	19	21
28	29	30	40	42

QUESTION 35

You are a Data Scientist working on a model that predicts fraudulent and non fraudulent transactions. You notice that 90% of the samples are non fraudulent which makes up the majority of the dataset. What are some methods you can use to address this issue in the data?

- ✓ Resample the dataset to correct imbalances of each transaction type
- ✗ Apply Principal Component Analysis (PCA) to undersample non fraudulent transactions
- ✗ Drop all fraudulent transaction before training the model
- ✗ Combine dataset with a public dataset that have a majority of fraudulent transactions — Selected
- ✗ Use K-means cluster to find outliers for non fraudulent transactions and use those as fraudulent samples
- ✓ Collect more data to even the imbalances in the dataset — Selected
- ✓ Try a different algorithm or a combination of algorithms on your dataset — Selected

EXPLANATION

There are many different ways to handle imbalanced data. What is important is that you have a keen sense for what damage an imbalanced dataset can cause and how to balance it. Handling Imbalanced Datasets - a review

Rate this question

80%

DOMAIN

Exploratory Data Analysis

2.1 Sanitize and prepare data for modeling 2.2 Perform feature engineering 2.3 Analyze and visualize data for machine learning Questions for this domain comprise **24% of the total questions** for this exam.

QUESTIONS

35	53	64	1	4
8	10	17	19	21
28	29	30	40	42

QUESTION 53

Your organization has the need to set up a petabyte scaled BI and dashboard analysis tool that will query millions of rows of data spread across thousands of files stored in S3. Your organization wants to save as much money as possible. Which solution will allows developers to run dozens if not hundreds or thousands of queries per day, and possibly scanning many TBs of data each, while still being cost effective?

- ✗ Data Pipeline and RDS
- ✗ AWS Glue Data Catalog and Amazon Athena — Selected
- ✓ EC2 Spot instances and Presto
- ✗ S3 Analytics

EXPLANATION

You pay a constant fee for the compute instances you are running (EC2 instances cost). The more machines you run and the bigger they are - the higher the fee, yes. But Presto is very efficient and if your data is correctly stored, a few commodity machines will do a great job if you are running your Presto cluster on the same region as your S3 bucket, and within one AZ, as there is no network or data transfer costs at all. The compute costs can be further optimized by using spot instances for worker nodes, and completely shutting them down off-hours (where applicable). Presto can deal with a lost worker node - which might slow down some queries but spot instances come at a great discount. Presto - Amazon EMR Presto | Distributed SQL Query Engine for Big Data

Rate this question

80%

DOMAIN

Exploratory Data Analysis

2.1 Sanitize and prepare data for modeling 2.2 Perform feature engineering 2.3 Analyze and visualize data for machine learning Questions for this domain comprise **24% of the total questions** for this exam.

QUESTIONS

35	53	64	1	4
8	10	17	19	21
28	29	30	40	42

QUESTION 64

During the data analysis portion of your machine learning process you have several hundred compressed JSON files stored in Amazon S3 around 200 MB in size. These files are categorised as semi-structured data and have already been crawled by AWS Glue to determine the schema associated with it. You have been using Amazon Athena to query your Amazon S3 data but finding it extremely expensive scanning 10 or more GBs of data each query. What are some techniques you can perform to cut down query execution costs?

- ✓ Only include columns in the queries being run that you need Selected
- ✗ Decompress and split files Selected
- ✓ Convert files to Apache Parquet or Apache ORC
- ✗ Break files into smaller files
- ✓ Partition your data Selected
- ✗ Convert files to CSV

EXPLANATION

Partitioning divides your table into parts and keeps the related data together based on column values such as date, country, region, etc. Partitions act as virtual columns. You define them at table creation, and they can help reduce the amount of data scanned per query, thereby improving performance. You can restrict the amount of data scanned by a query by specifying filters based on the partition. Compressing your data can speed up your queries significantly, as long as the files are either of an optimal size (see the next section), or the files are splittable. The smaller data sizes reduce network traffic from Amazon S3 to Athena. Splittable files allow the execution engine in Athena to split the reading of a file by multiple readers to increase parallelism. If you have a single unsplittable file, then only a single reader can read the file while all other readers sit idle. Not all compression algorithms are splittable. The following table lists common compression formats and their attributes. Queries run more efficiently when reading data can be parallelized and when blocks of data can be read sequentially. Ensuring that your file formats are splittable helps with parallelism regardless of how large your files may be. However, if your files are too small (generally less than 128 MB), the execution engine might be spending additional time with the overhead of opening Amazon S3 files, listing directories, getting object metadata, setting up data transfer, reading file headers, reading compression dictionaries, and so on. On the other hand, if your file is not splittable and the files are too large, the query processing waits until a single reader has completed reading the entire file. That can reduce parallelism. Apache Parquet and Apache ORC are popular columnar data stores. They provide features that store data efficiently by employing column-wise compression, different encoding, compression based on data type, and predicate pushdown. They are also splittable. Generally, better compression ratios or skipping blocks of data means reading fewer bytes from Amazon S3, leading to better query performance. When running your queries, limit the final SELECT statement to only the columns that you need instead of selecting all columns. Trimming the number of columns reduces the amount of data that needs to be processed through the entire query execution pipeline. This especially helps when you are querying tables that have large numbers of columns that are string-based, and when you perform multiple joins or aggregations. Top 10 Performance Tuning Tips for Amazon Athena | AWS Big Data Blog

Rate this question

83%

DOMAIN

Modeling

3.1 Frame business problems as machine learning problems 3.2 Select the appropriate model(s) for a given machine learning problem 3.3 Train machine learning models 3.4 Perform hyperparameter optimization 3.5 Evaluate machine learning models Questions for this domain comprise **36% of the total questions** for this exam.

QUESTIONS

48	50	55	63	2
5	6	9	11	13
14	18	22	23	26
27	31	32	38	43
54	56	58	61	

QUESTION 48

Which of the following is NOT a valid use-case for incremental training?

- ✗ Train several variants of a model, either with different hyperparameter settings or using different datasets.
- ✓ Rebuilt model artifacts which you have accidentally deleted.
- ✗ Train a new model using an expanded dataset that contains an underlying pattern that was not accounted for in the previous training and which resulted in poor model performance. Selected
- ✗ Resume a training job that was stopped.
- ✗ Use the model artifacts or a portion of the model artifacts from a popular publicly available model in a training job. You don't need to train a new model from scratch.
- ✗ Train several variants of a model, either with different hyperparameter settings or using different datasets.

EXPLANATION

Incremental training is used for all of these except rebuilding model artifacts. Incremental training picks up where another training job left off by using the existing artifacts created by the prior training job. Incremental Training in Amazon SageMaker - Amazon SageMaker

Rate this question

83%

DOMAIN

Modeling

3.1 Frame business problems as machine learning problems 3.2 Select the appropriate model(s) for a given machine learning problem 3.3 Train machine learning models 3.4 Perform hyperparameter optimization 3.5 Evaluate machine learning models Questions for this domain comprise **36% of the total questions** for this exam.

QUESTIONS

48	50	55	63	2
5	6	9	11	13
14	18	22	23	26
27	31	32	38	43
54	56	58	61	

QUESTION 50

A ski equipment company is trying to predict the expected sales from a line of ski goggles. They have never sold this type of product before, but they do have some historic sales data for other products which they believe have similar market adoption curves. What would be your first algorithm of choice among the built-in SageMaker algorithms for this use-case?

- ✗ Factorization Machines
- ✓ DeepAR
- ✗ Linear Learner
- ✗ K-Nearest Neighbor Selected
- ✗ XGBoost

EXPLANATION

While there may be many ways to create a forecasting model, the SageMaker DeepAR Forecasting algorithm is most closely positioned for this use case. We can use multiple sets of historic data together to create a more refined forecast than if we were just using a single product sales history dataset. DeepAR Forecasting Algorithm - Amazon SageMaker

Rate this question

83%

DOMAIN

Modeling

3.1 Frame business problems as machine learning problems 3.2 Select the appropriate model(s) for a given machine learning problem 3.3 Train machine learning models 3.4 Perform hyperparameter optimization 3.5 Evaluate machine learning models Questions for this domain comprise **36% of the total questions** for this exam.

QUESTIONS

48	50	55	63	2
5	6	9	11	13
14	18	22	23	26
27	31	32	38	43
54	56	58	61	

QUESTION 55

When evaluating a model after the training and testing process, you notice that the error rate during training is high but the error rate during testing is low. Which of the following could be the reason for obtaining these error rates?

- ✗ You should train for a longer period of time. Selected
- ✓ You have a data issue with both your training and testing datasets. Selected
- ✗ Your model is overfitting the training data.
- ✓ You have a programmatic issue with your algorithm.
- ✗ Your model is underfitting the testing data.
- ✗ You need to re-evaluate the section of your algorithm.

EXPLANATION

When training error is high and testing error is low, this is highly unusual as it infers that the model is somehow predicting better than the data which was used to train the model. This is usually an indicator of a data issue or some systemic problem in the algorithm. Overfitting - Wikipedia

Rate this question

83%

DOMAIN

Modeling

3.1 Frame business problems as machine learning problems 3.2 Select the appropriate model(s) for a given machine learning problem 3.3 Train machine learning models 3.4 Perform hyperparameter optimization 3.5 Evaluate machine learning models Questions for this domain comprise **36% of the total questions** for this exam.

QUESTIONS

48	50	55	63	2
5	6	9	11	13
14	18	22	23	26
27	31	32	38	43
54	56	58	61	

QUESTION 63

You are trying to classify a number of items based on different features into one of 6 groups (books, electronics, movies, etc.) based on features. Which algorithm would be best suited for this type of problem?

- ✗ Use Linear Learner with predictor set to regressor
- ✗ Use a stochastic approach when choosing target parameters and recommended ranges
- ✗ Use regression forest with the number of trees set to the number of categories
- ✗ Use K-Means algorithm with k set to the number of classes Selected
- ✓ Use XGBoost with objective set to multi:softmax

EXPLANATION

Both XGBoost and Linear Learner are perfect choices for multi classification problems. When we are trying to solve a multi classification problem using XGBoost we set the objective hyperparameter to multi:softmax and when using the Linear Learner algorithm, we set the predictor hyperparameter to multiclass_classifier. XGBoost Algorithm - Amazon SageMaker

Rate this question

62%

DOMAIN

Data Engineering

1.1 Create data repositories for machine learning 1.2 Identify and implement a data-ingestion solution 1.3 Identify and implement a data-transformation solution Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

15 16 37 44 65
3 20 36 39 45
46 57 62

QUESTION 15

You are a data scientist that has been tasked with setting up an Amazon Elastic Map Reduce (EMR) cluster to host your organization's data lake. You also need to setup this cluster for machine learning processes and it has been decided to use Amazon SageMaker libraries as the machine learning platform. What steps do you need to take to start using SageMaker with your EMR cluster data lake?

- ✓ Download the aws-sagemaker-spark-sdk component along with Spark on your EMR cluster — Selected
- ✓ Run your SageMaker Spark application on EMR by submitting your Spark application jar and any additional dependencies your Spark application uses
- ✗ Convert EMR DataFrame to CSV and use that to train and infer your model
- ✗ Use Apache Mahout within an EMR Notebook to train and infer your model
- ✗ Ensure the EMR cluster and SageMaker hosted model are in the same region to make successful inferences — Selected

EXPLANATION

SageMaker Spark is an open source Spark library for Amazon SageMaker. With SageMaker Spark you can construct Spark ML Pipelines using Amazon SageMaker stages. These pipelines interleave native Spark ML stages and stages that interact with SageMaker training and model hosting. With SageMaker Spark, you can train on Amazon SageMaker from Spark DataFrames using Amazon-provided ML algorithms or using your own algorithms -- all at Spark scale. SageMaker Spark README.md Adding a Spark Step - Amazon EMR

Rate this question

62%

DOMAIN

Data Engineering

1.1 Create data repositories for machine learning 1.2 Identify and implement a data-ingestion solution 1.3 Identify and implement a data-transformation solution Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

15 16 37 44 65
3 20 36 39 45
46 57 62

QUESTION 16

You are working for an online shopping platform that records actions made by its users. This information is captured in multiple JSON files stored in S3. You have been tasked with moving this data into Amazon Redshift database tables as part of a data lake migration process. Which of the following needs to occur to achieve this in the most efficient way?

- ✓ Launch an Amazon Redshift cluster and create database tables. — Selected
- ✗ Setup DynamoDB table and use Data Pipeline to load the S3 data into DynamoDB table.
- ✗ Use the INSERT command to load the tables from the data files on Amazon S3.
- ✗ Use multiple concurrent COPY commands to load the table from each JSON file. — Selected
- ✓ Use COPY commands to load the tables from the data files on Amazon S3. — Selected
- ✗ Use COPY commands to load the tables from the data files on DynamoDB.
- ✓ Troubleshoot load errors and modify your COPY commands to correct the errors.

EXPLANATION

You can add data to your Amazon Redshift tables either by using an INSERT command or by using a COPY command. At the scale and speed of an Amazon Redshift data warehouse, the COPY command is many times faster and more efficient than INSERT commands. You can load data from an Amazon DynamoDB table, or from files on Amazon S3, Amazon EMR, or any remote host through a Secure Shell (SSH) connection. When loading data from S3, you can load table data from a single file, or you can split the data for each table into multiple files. The COPY command can load data from multiple files in parallel. Using a COPY Command to Load Data - Amazon Redshift

Rate this question

62%

DOMAIN

Data Engineering

1.1 Create data repositories for machine learning 1.2 Identify and implement a data-ingestion solution 1.3 Identify and implement a data-transformation solution Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

15	16	37	44	65
3	20	36	39	45
46	57	62		

QUESTION 37

You work for a company that builds custom python libraries for transforming and preprocessing data sets before using them in BI tools and machine learning pipelines. One of your customers is using your libraries and has the need to include them within their AWS Glue pipeline. What suggestions can you make to allow your customers to use your libraries within their AWS Glue pipelines?

- ✗ AWS Glue does not support custom code outside of PySpark and Scala implemented libraries — Selected
- ✗ Give the custom code to the customer allowing the customers to upload the code onto AWS Glue and use within an ETL job.
- ✓ Upload the custom library as a .zip archive onto S3. Before your customers create an ETL job, include the S3 link as a script library and job parameter
- ✗ Upload all of the library files onto S3. Before your customers create an ETL job, include the S3 link as a script library and job parameter

EXPLANATION

When you are creating a new Job on the console, you can specify one or more library .zip files by choosing Script Libraries and job parameters (optional) and entering the full Amazon S3 library path(s) in the same way you would when creating a development endpoint: Using Python Libraries with AWS Glue - AWS Glue

Rate this question

62%

DOMAIN

Data Engineering

1.1 Create data repositories for machine learning 1.2 Identify and implement a data-ingestion solution 1.3 Identify and implement a data-transformation solution Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

15	16	37	44	65
3	20	36	39	45
46	57	62		

QUESTION 44

You have been tasked with collecting 100-byte events from hundreds or thousands of low power devices and writing records into a Kinesis stream. You have Amazon Elastic Compute Cloud (EC2) instances serving as a proxy for these events. You must add logic for batching or multithreading, in addition to retry logic and record de aggregation at the consumer side. Which service can you use to handle all of this for you?

- ✓ Using the Kinesis Producer Library (KPL)
- ✗ Using a combination of SQS and Lambda for retry logic and batching respectively.
- ✗ Using the APIs for Kinesis Streams
- ✗ Using the Kinesis Client Library (KCL) — Selected
- ✗ Using EMR cluster as intermediate logic mechanism
- ✗ Using the Amazon Kinesis Agent

EXPLANATION

KPL simplifies producer application development, allowing developers to achieve high write throughput to one or more Kinesis streams. The KPL is an easy-to-use, highly configurable library that you install on your hosts that generate the data that you wish to stream to Kinesis Streams. It acts as an intermediary between your producer application code and the Kinesis Streams API actions. The KPL performs the following primary tasks: 1) Writes to one or more Kinesis streams with an automatic and configurable retry mechanism 2) Collects records and uses PutRecords to write multiple records to multiple shards per request 3) Aggregates user records to increase payload size and improve throughput 4) Integrates seamlessly with the Amazon Kinesis Client Library (KCL) to de-aggregate batched records on the consumer 5) Submits Amazon CloudWatch metrics on your behalf to provide visibility into producer performance Streaming Data Solutions on AWS with Amazon Kinesis

62%

DOMAIN

Data Engineering

1.1 Create data repositories for machine learning 1.2 Identify and implement a data-ingestion solution 1.3 Identify and implement a data-transformation solution Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

15	16	37	44	65
3	20	36	39	45
46	57	62		

QUESTION 65

You are working for a hot new startup that calculates different metrics about their customers depending on how much money they spend on a weekly, quarterly, and yearly basis. These metrics are classified as elite, novice, and beginner. Depending on their ranking they get more/less discounts and placed in higher/lower priority for customer support. Your machine learning model should take this ordering into consideration. The algorithm you have chosen expects all numerical inputs. What can be done to handle these classification values?

- ✗ Apply random numbers to each classification value and apply gradient descent until the values converge to expect results
- ✗ Use one-hot encoding techniques to map values for each classification dropping the original classification feature
- ✓ Experiment with mapping different values for each status and see which works best
- ✗ Use one-hot encoding techniques to map values for each classification Selected

EXPLANATION

Since these classification values are ordinal (order does matter) we cannot use one-hot encoding techniques. We either need to map these values to a scale, or we train our model with different encodings and seeing which encoding works best. Ordinal data - Wikipedia

Rate this question

69%

DOMAIN

Machine Learning Implementation and Operations

4.1 Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance 4.2 Recommend and implement the appropriate machine learning services and features for a given problem 4.3 Apply basic AWS security practices to machine learning solutions 4.4 Deploy and operationalize machine learning solutions Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

12	33	41	47	7
24	25	34	49	51
52	59	60		

QUESTION 12

After several weeks of working on a model for genome mapping, you believe you have perfected it and now want to deploy it to a platform that will provide the highest performance. Which of the following AWS platforms will provide the highest performance for this compute-intensive model?

- ✗ EC2 P2 Instance Selected
- ✗ EC2 X1 Instance
- ✓ EC2 F1 instance
- ✗ EC2 G3 Instance
- ✗ EC2 M2 Instance

EXPLANATION

Of these instance types, F1 instance types provide the most performance via use of FPGAs to enable delivery of custom hardware accelerations. Good target applications for F1 are ones that have a modest number of distinct operations that account for significant portions of application run-time. Examples of such applications include big data analytics, genomics, electronic design automation (EDA), image and video processing, compression, security, and search/analytics. The down-side of FPGAs is that they usually require very specialized knowledge and programming to deploy models on them.
Amazon EC2 F1 Instances

Rate this question

69%

DOMAIN

Machine Learning Implementation and Operations

4.1 Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance 4.2 Recommend and implement the appropriate machine learning services and features for a given problem 4.3 Apply basic AWS security practices to machine learning solutions 4.4 Deploy and operationalize machine learning solutions Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

12 33 41 47 7
24 25 34 49 51
52 59 60

QUESTION 33

You have been tasked with using Polly to translate text to speech in the company announcements that launch weekly. The problem you are encountering is how Polly is incorrectly translating the companies acronyms. What can be done for future tasks to help prevent this?

- ✗ Use Amazon Comprehend to pull parts of speech and use to help pronounce acronyms
- ✗ Use speech marks for input text documents — Selected
- ✗ Use Amazon Transcribe to first map the acronyms to pronunciations then include them in the Amazon polly pipeline
- ✓ Create dictionary lexicon — Selected
- ✓ Use SSML tags in documents

EXPLANATION

Using SSML-enhanced input text gives you additional control over how Amazon Polly generates speech from the text you provide. Using these tags allows you to substitute a different word (or pronunciation) for selected text such as an acronym or abbreviation. You can also create a dictionary lexicon to apply to any future tasks instead of apply SSML to each individual document. Amazon Polly

Rate this question

69%

DOMAIN

Machine Learning Implementation and Operations

4.1 Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance 4.2 Recommend and implement the appropriate machine learning services and features for a given problem 4.3 Apply basic AWS security practices to machine learning solutions 4.4 Deploy and operationalize machine learning solutions Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

12 33 41 47 7
24 25 34 49 51
52 59 60

QUESTION 41

You have setup a group of SageMaker Notebook instances for your company's data scientists. You wanted to uphold your company's philosophy on least privilege and disabled Internet access for the notebooks. However, the data scientists report that they are unable to import certain key libraries from the Internet into their notebooks. What is the most efficient path?

- ✗ Create a series of EC2 instances outside of the VPC and install Jupyter Notebook on those instances. Have the scientists use those instances instead of SageMaker.
- ✗ Advise the data scientists that it is not possible to import libraries from the internet given the company's least privilege philosophy. — Selected
- ✗ Suggest that the scientists choose different libraries that are open source and do not pose a threat to company policy.
- ✓ Create a NAT gateway within the Notebook VPC and associated default route to the NAT gateway.
- ✗ Create a VPC Gateway Endpoint that bridges between the VPC and the desired Internet location of the required libraries.

EXPLANATION

Notebook instances are Internet-enabled by default but you can disable this internet access. If you do so and you still need to access the Internet from the Notebook instances, you must create a NAT gateway, appropriate route and security groups that will allow outbound connections to the Internet. Notebook Instance Security - Amazon SageMaker

Rate this question

69%

DOMAIN

Machine Learning Implementation and Operations

4.1 Build machine learning solutions for performance, availability, scalability, resiliency, and fault tolerance 4.2 Recommend and implement the appropriate machine learning services and features for a given problem 4.3 Apply basic AWS security practices to machine learning solutions 4.4 Deploy and operationalize machine learning solutions Questions for this domain comprise **20% of the total questions** for this exam.

QUESTIONS

12 33 41 47 7
24 25 34 49 51
52 59 60

QUESTION 47

You have setup autoscaling for your deployed model using SageMaker Hosting Services. You notice that in times of heavy load spikes, it takes a long time for the hosted model to scale out in response to the load. How might you increase the reaction time of auto-scaling?

- ✗ Disable CloudWatch advanced tracking metrics.
- ✗ Change the scale metric from InvocationsPerInstance to MemoryUtilization. Selected
- ✗ Create a new target metric based on time since last scale event.
- ✓ Reduce the cooldown period for automatic scaling.
- ✗ Change the timeout in the auto-scaling Lambda function.

EXPLANATION

When scaling responsiveness is not as fast as you would like, you should look at the cooldown period. The cooldown period is a duration when scale events will be ignored, allowing the new instances to become established and take on load. Decreasing this value will launch new variant instance faster. Automatically Scale Amazon SageMaker Models - Amazon SageMaker

Rate this question