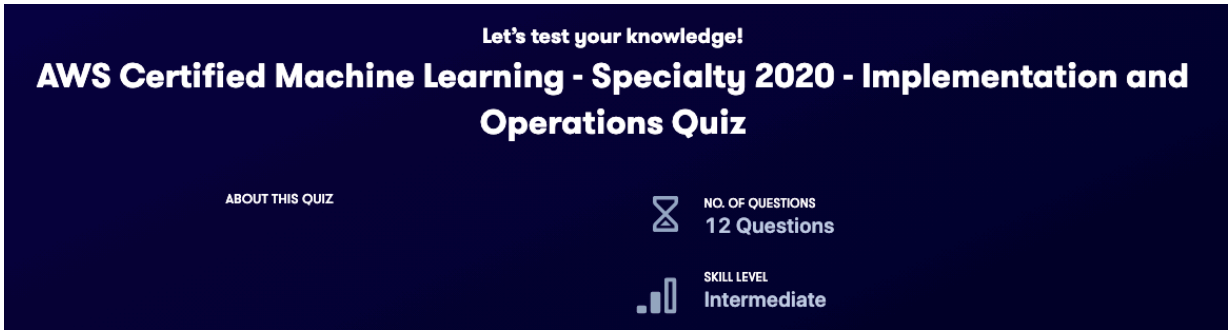
Cloud Guru - 8 - Implementation and Operations Quiz



QUESTION 1

You are helping a digital asset media company create a system which can automatically extract metadata from photographs submitted by freelance photographers. They want a solution that is robust, cost-effective and flexible but they don't want to manage lots of infrastructure. What would you recommend?

- [x] Make use of Amazon Rekognition for metadata extraction.
- [ ] Make use of Amazon Comprehend to extract metadata from the images.
- [ ] Build a model using the Semantic Segmentation algorithm and host it using SageMaker Hosting Services.
- [ ] Build a model using Image Analysis to extract metadata from images and host it using Lambda and the API Gateway.
- [ ] Build a model using Object Detection to extract metadata from images and host it using EC2.

Good work!

The option that provides the ability to extract metadata from images with the lowest maintenance requirements would be Amazon Rekognition.

QUESTION 2

You need to chain together three different algorithms for a model you are creating. You need to run PCA, RCF, and LDA in succession. What is the recommended way to do this?

- [x] Use an Inference Pipeline to link together these algorithms.
- [ ] Use AWS Batch to create a script that will trigger each algorithm in sequence.
- [ ] Use Lambda Step Functions to link together the separate training jobs.
- [ ] You cannot run SageMaker built-in algorithms together. You will need to create individual training jobs and manually execute them via SDK or Console.

Good work!

An inference pipeline is an Amazon SageMaker model composed of a linear sequence of two to five containers that process requests for inferences on data. Amazon SageMaker Inference Pipelines enable the definition and deployment of any combination of pretrained Amazon SageMaker built-in algorithms and your own custom algorithms packaged in Docker containers.

QUESTION 3

You need to increase the performance of your Image Classification inference endpoint and want to do so in the most cost-effective manner. What should you choose?

- Create a new endpoint deployment that uses a single-CPU instance given the algorithm being used.
- Create a new production variant that uses a multi-GPU instance.
- ✓ Redeploy the endpoint using Elastic Inference added to the production variant.
- Offload some traffic to a less costly AWS region.
- Create an additional production variant which is the same as the original variant and direct 50% of the traffic to that variant.

Good work!

By using Amazon Elastic Inference (EI), you can speed up the throughput and decrease the latency of getting real-time inferences from your deep learning models that are deployed as Amazon SageMaker hosted models, but at a fraction of the cost of using a GPU instance for your endpoint.

QUESTION 4

Your company has just established a policy that says all data must be encrypted at rest. You are currently using SageMaker to host Jupyter Notebook instances for your data scientists. What is the most direct path for you to ensure you are compliant?

- Create an IAM resource policy that requires encryption of all data.
- Recreate the Notebook Instance and select an encryption key from Amazon Certificate Manager.
- Migrate the Notebooks into CodeCommit and redeploy the Notebook instances on-prem using encrypted storage.
- Require the data scientists to use Amazon VPN when connecting to their Notebook instances.
- ✓ Recreate the Notebook Instances and select an encryption key from KMS.
- Create an EC2 instance using local volume encryption then migrate over the existing Jupyter Notebooks.

Good work!

Because SageMaker Notebook Instances directly support KMS keys, the most direct path would be to recreate the notebook instances with a KMS key selected.

QUESTION 5

You are helping a client design a landscape for their mission critical ML model based on DeepAR deployed using SageMaker Hosting Services. Which of the following would you recommend they do to ensure high availability?

- [x] Ensure that InitialInstanceCount is at least 2 or more in the endpoint production variant.
- [ ] Include Elastic Inference in the endpoint configuration.
- [ ] Keep a copy of all the DeepAR code in a Glacier Vault for safekeeping.
- [ ] Create a duplicate endpoint in another region using Amazon Forecast.
- [ ] Recommend that they deploy using EKS in addition to the SageMaker Hosting deployment.

Good work!

AWS recommends that customers deploy a minimum of 2 instances for mission critical workloads when using SageMaker Hosting Services. SageMaker will automatically spread multiple instances across different AZs within a region.

QUESTION 6

You are preparing to release an updated version of your latest machine learning model. It is provided to about 3,000 customers who use it in a SaaS capacity. You want to minimize customer disruption, minimize risk and be sure the new model is stable before full deployment. What is the best course of action?

- Use a canary deployment to evaluate the new model then distribute the new production URL to your customers.
- ✓ Conduct an A/B test first then use a phased rollout.
- Use a continuous integration process to preserve the stability of the new model and deploy in a "Big Bang" manner.
- Perform offline validation then cut over all at once to the new version to minimize risk.

Good work!

Of the options provided, using an A/B test to first evaluate the stability of the update with a small group, then using a phased rollout, seems to fulfill the objectives. A phased rollout could be done as an extension of the A/B test by simply moving 100% of the traffic to the B version.

QUESTION 7

You want to deploy an XGBoost-backed model to a fleet of traffic sensors using Raspberry Pis as the local compute component. Will this work?

- No, best practice says that you should not deploy ML models into the field but rather use a centralized inference landscape.
- No, a Raspberry Pi is not powerful enough to run an ML model using XGBoost.
- ✓ Yes, you can use SageMaker Neo to compile the model into a format that is optimized for the ARM processor on the Raspberry Pi.
- No, XGBoost cannot be compiled to run on an ARM processor. It can only run on x86 architectures.
- Yes, you can deploy the model using Amazon Robomaker using the native ARM support.

Good work!

SageMaker Neo provides a way to compile XGBoost models, which are optimized for the ARM processor in the Raspberry Pi.

QUESTION 8

To make use of your published model in a custom application, what must you do?

- Instruct SageMaker to generate a unique endpoint URL for your application.
- Use a Lambda function to perform the inferences for your application.
- ✓ Use the SageMaker API InvokeEndpoint() method via SDK.
- Use the CloudTrail API to monitor for inference requests and trigger the SageMaker model endpoint.
- Create an entry in Route 53 to point your desired DNS name to the endpoint.

Good work!

To use your endpoint for inferences, you can use the SageMaker API InvokeEndpoint() method with the automatically generated HTTPS URL specified.

QUESTION 9

You have decided to use SageMaker Hosting Services to deploy your newly created model. What is the next required set after you have created your model?

- Turn on CloudWatch logging for your model.
- Nothing additional is required. SageMaker Hosting Services is enabled with every model created on SageMaker.
- ✓ Create an endpoint configuration.
- Configure Route 53 to point your desired DNS name to the endpoint.
- Create an endpoint.

Good work!

Once you have a model in SageMaker you must next create endpoint configurations, which include the specific model to use, the instance information, and potentially the initial weight if you are deploying multiple variants.

QUESTION 10

Your company has just discovered a security breach occurred in a division separate from yours but has ordered a full review of all access logs. You have been asked to provide the last 180 days of access to the three SageMaker Hosted Service models that you manage. When you set up these deployments, you left everything default. How will you be able to respond?

- [x] Use CloudTrail to pull a list of all access to the models for the last 90 days. Any data beyond 90 days is unavailable.
- [ ] Use SageMaker Detailed Logging to produce a CSV file of access from the past 180 days.
- [ ] Use CloudTrail to pull a list of all access to the ML models for the last 180 days.
- [ ] Use CloudWatch along with IPInsights to analyse the logs for suspicious activity from the past 180 days then download these records.
- [ ] Use CloudWatch to pull a list of all access records for the ML models. Make use of a Python library to parse out only the access records.

Good work!

CloudTrail is the proper service if you want to see who has sent API calls to your SageMaker Hosted model but, by default, it will only store the last 90 days of events. You can configure CloudTrail to store an unlimited amount of logs on S3 but this is not turned on by default. Whilst CloudTrail is not necessarily an Access Log, it performs the same auditing functions you might expect; and an auditor may not necessarily be familiar with the nuances of AWS

QUESTION 11

You have been asked to build an automated chatbot for customer service. If the initial interaction with the customer seems negative or the customer is upset or unhappy, you want to immediately transfer that chat session over to a live human. What is the simplest way to implement this feature?

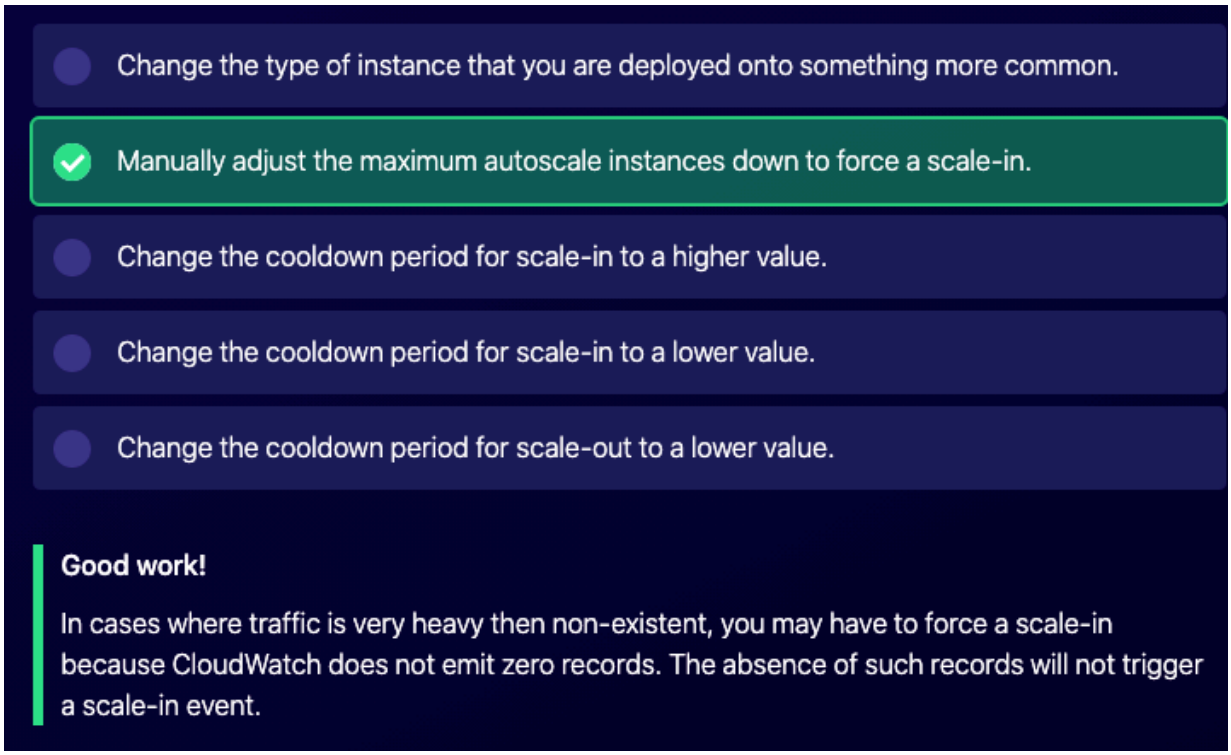
- [x] Use Amazon Lex to take in the customer's initial comments, then process them through Amazon Comprehend to determine sentiment. If sentiment is negative, hand the chat session over to a live customer support person.
- [ ] Use Amazon Comprehend to take in the customer's initial comments, then process them through Amazon Personalize to determine sentiment. If sentiment is negative, hand the chat session over to a live customer support person.
- [ ] Use LDA to create an NLP model that can understand the sentiment of the customer's comments. Create a Lambda function to redirect the chat session over to a live customer support person.
- [ ] Use XGBoost to create a binary classification model to decide if a customer's initial comments are negative or positive. If negative, redirect the chat session over to a live customer support person.
- [ ] Use IPInsights to identify the customer by their IP address. If they have had a recent bad experience as logged in the CRM system, direct them to a live customer support person.

Good work!

While this use-case could be done using SageMaker algorithms, the simplest way is to use Amazon Lex for NLP and Amazon Comprehend for sentiment analysis.

QUESTION 12

Your newly deployed model gets heavy usage on Monday then no usage the rest of the week. To accommodate this heavy usage, you make use of auto-scaling to adjust to the inbound request load. After several weeks in production, you notice a large number of scaled resources going unused and thus consuming money for no good reason. What might you do to resolve this?


Change the type of instance that you are deployed onto something more common.
Manually adjust the maximum autoscale instances down to force a scale-in.
Change the cooldown period for scale-in to a higher value.
Change the cooldown period for scale-in to a lower value.
Change the cooldown period for scale-out to a lower value.
Good work!
In cases where traffic is very heavy then non-existent, you may have to force a scale-in because CloudWatch does not emit zero records. The absence of such records will not trigger a scale-in event.


100%
Congratulations!
You passed AWS Certified Machine Learning - Specialty 2020 - Implementation and Operations Quiz!