Production Machine Learning Systems

Quiz Questions and Answers

In each quiz, the questions are randomized, so when you take the quiz, the questions may be in a different order than what you see here.

Architecting Production ML Systems

Question 1
What percent of system code does the ML model account for?

*A: 5%
Feedback: Correct!

B: 25%
Feedback: This answer is not correct. Review the module.

C: 50%
Feedback: This answer is not correct. Review the module.

D: 90%
Feedback: This answer is not correct. Review the module.

Question 2
Match the three types of data ingest with an appropriate source of training data.

*A: Streaming (Pub/Sub), structured batch (BigQuery), unstructured batch (Cloud Storage)
Feedback: Correct! On Google Cloud, the three types of data ingestion map to three different products. If you are ingesting streaming data, use Pub/Sub. If you are ingesting structured data directly into your ML model, use BigQuery, and if you are transforming data from training so that you can train on it later, read from Cloud Storage.

B: Streaming (BigQuery), structured batch (Pub/Sub), unstructured batch (Cloud Storage)
Feedback: This answer is not correct. You wouldn't ingest streaming data from BigQuery, although you could stream to it. Pub/Sub is a poor place to store your batch data, although

you might use it to replay events.

C: Streaming batch (Dataflow), structured batch (BigQuery), stochastic (App Engine)
Feedback: This answer is not correct. These are just made up terms.

Question 3
What is the responsibility of model evaluation and validation components?

*A: To ensure that the models are good before moving them into a production/staging environment.
Feedback: Correct!

B: To ensure that the models are not good before moving them into a staging environment.
Feedback: This answer is not correct. Review the module.

C: To ensure that the models are good after moving them into a production/staging environment.
Feedback: This answer is not correct. Review the module.

D: To ensure that the models are not good after moving them into a staging environment.
Feedback: This answer is not correct. Review the module.

Question 4
Which type of training do you use if your data set doesn't change over time?

*A: Static training
Feedback: Correct!

B: Dynamic training
Feedback: This answer is not correct. Review the module.

C: Real-time training
Feedback: This answer is not correct. Review the module.

D: Online training
Feedback: This answer is not correct. Review the module.

Question 5
Which type of logging should be enabled in the online prediction that logs the stderr and stdout streams from your prediction nodes to Cloud Logging and can be useful for debugging?

A: Access logging
Feedback: This answer is not correct. Review the module.

B: Request-response logging
Feedback: This answer is not correct. Review the module.

*C: Container logging
Feedback: Correct!

D: Cloud logging
Feedback: This answer is not correct. Review the module.

Question 6
Vertex AI has a unified data preparation tool that supports image, tabular, text, and video content. Where are uploaded datasets stored in Vertex AI?

A: A Google Cloud database that acts as an output for both AutoML and custom training jobs.
Feedback: This answer is not correct. Review the module.

B: A Google Cloud Storage bucket that acts as an output for both AutoML, custom training jobs, serialized training jobs.
Feedback: This answer is not correct. Review the module.

*C: A Google Cloud Storage bucket that acts as an input for both AutoML and custom training jobs.
Feedback: Correct!

D: A Google Cloud database that acts as an input for both AutoML and custom training jobs.
Feedback: This answer is not correct. Review the module.

Question 7
In the featurestore, the timestamps are an attribute of the feature values, not a separate resource type.

*A: True
Feedback: Correct!

B: False
Feedback: This answer is not correct. Review the module.

Question 8
When you use the data to train a model, Vertex AI examines the source data type and feature values and infers how it will use that feature in model training. This is called the _____for that feature.

A. Duplication
Feedback: This answer is not correct. Review the module.

B. Transmutation
Feedback: This answer is not correct. Review the module.

C. Translation
Feedback: This answer is not correct. Review the module.

*D. Transformation
Feedback: Correct!

Designing Adaptable ML Systems

Question 1
Which of the following models are susceptible to a feedback loop? Check all that apply.

*A: A traffic-forecasting model that predicts congestion at highway exits near the beach, using beach crowd size as one of its features.
Feedback: Correct! Some beachgoers are likely to base their plans on the traffic forecast. If there is a large beach crowd and traffic is forecast to be heavy, many people may make alternative plans. This may depress beach turnout, resulting in a lighter traffic forecast, which then may increase attendance, and the cycle repeats.

*B: A university-ranking model that rates schools in part by their selectivity (the percentage of students who applied that were admitted).
Feedback: Correct! The model's rankings may drive additional interest to top-rated schools, increasing the number of applications they receive. If these schools continue to admit the same number of students, selectivity will increase (the percentage of students admitted will go down). This will boost these schools' rankings, which will further increase prospective student interest, and so on...

*C: A book-recommendation model that suggests novels its users may like based on their popularity (i.e., the number of times the books have been purchased).
Feedback: Correct! Book recommendations are likely to drive purchases, and these additional sales will be fed back into the model as input, making it more likely to recommend these same books in the future.

D: A face-attributes model that detects whether a person is smiling in a photo, which is regularly trained on a database of stock photography that is automatically updated monthly.
Feedback: This answer is not correct. There is no feedback loop here, because model predictions don't have any impact on the photo database. However, versioning of the input data is a concern here, because these monthly updates could potentially have unforeseen effects on the model.

E: A housing-value model that predicts house prices, using size (area in square meters), number of bedrooms, and geographic location as features.
Feedback: This answer is not correct. A house's location, size, or number of bedrooms cannot be quickly changed in response to price forecasts, which makes a feedback loop unlikely. However, there is potentially a correlation between size and number of bedrooms (larger homes are likely to have more rooms) that may need to be analyzed.

F: An election-results model that forecasts the winner of a mayoral race by surveying 2% of voters after the polls have closed.
Feedback: This answer is not correct. If the model does not publish its forecast until after the polls have closed, its predictions cannot affect voter behavior.

Question 2
Suppose you are building an ML-based system to predict the likelihood that a customer will leave a positive review. The user interface that customers leave reviews on changed a few months ago, but you don't know about this. Which of these is a potential consequence of mismanaging this data dependency?

*A: Losses in prediction quality
Feedback: Correct! For example, a review might be easier to write, and so your prediction of whether someone will leave a review (whether good or bad) is too low because it was trained on reviews that resulted from the older, harder-to-use user interface

B: Change in model serving signature
Feedback: This answer is not correct. Your model structure doesn't change just because it's easier or harder to leave reviews.

C: Change in ability of model to be part of a streaming ingest
Feedback: This answer is not correct. Your model structure doesn't change just because it's easier or harder to leave reviews.

Question 3
What is training skew caused by?
*A: Your development and production environments are different, or different code is used in the training environment than in the development environment.
Feedback: Correct! Different versions may cause predictions to be significantly slower or consume more memory in the training environment than in the development environment. Different codes may result in different performance.

B: The prediction environment is slower than the training environment. Feedback: This answer is not correct. Training may take longer in development than in production, but the training is the same.

C: The Cloud Storage you load your data from in the training environment is physically closer than the Cloud Storage you load your data from in the production environment. Feedback:

This answer is not correct. The distance of where the data is stored to the processing device does not impact prediction performance.

D: Starting and stopping of the processing when training the model.
Feedback: This answer is not correct. Starting and stopping the processing makes no difference to the training.

Question 4
Gradual drift is used for which of the following?

A: A new concept that occurs within a short time
Feedback: This answer is not correct. Review the module.

B: A new concept that rapidly replaces an old one over a short period of time
Feedback: This answer is not correct. Review the module.

*C: An old concept that incrementally changes to a new concept over a period of time
Feedback: Correct!

D: An old concept that may reoccur after some time
Feedback: This answer is not correct. Review the module.

Question 5
Which of the following tools help software users manage dependency issues?

A: Modular programs
Feedback: This answer is not correct. Review the module.

*B: Maven, Gradle, and Pip
Feedback: Correct!

C: Monolithic programs
Feedback: This answer is not correct. Review the module.

D: Polylithic programs
Feedback: This answer is not correct. Review the module.

Question 6
Which component identifies anomalies in training and serving data and can automatically create a schema by examining the data?

A: Data ingestion
Feedback: This answer is not correct. Review the module.

B: Data transform
Feedback: This answer is not correct. Review the module.

*C: Data validation
Feedback: Correct!

D: Data identifier
Feedback: This answer is not correct. Review the module.

Question 7
What is the shift in the actual relationship between the model inputs and the output called?

A: Data drift
Feedback: This answer is not correct. Review the module.

*B: Concept drift
Feedback: Correct!

C: Prediction drift
Feedback: This answer is not correct. Review the module.

D: Label drift
Feedback: This answer is not correct. Review the module.

Designing High-Performance ML Systems

Question 1
If each of your examples is large in terms of size and requires parsing, and your model is relatively simple and shallow, your model is likely to be:

*A: I/O bound, so you should look for ways to store data more efficiently and ways to parallelize the reads.
Feedback: Correct! Your ML training will be I/O bound if the number of inputs is large or heterogeneous (requires parsing) or if the model is so small that the compute requirements are trivial. This also tends to be the case if the input data is on a storage system with low throughput. If you are I/O bound, look at storing the data more efficiently, storing the data on a storage system with higher throughput, or parallelizing the reads. Although it is not ideal, you might consider reducing the batch size so that you are reading less data in each step.

B: CPU-bound, so you should use GPUs or TPUs.
Feedback: This answer is not correct. This doesn't sound like computational power is your limiting factor.

C: Latency-bound, so you should use faster hardware
Feedback: This answer is not correct. Review I/O-bound, CPU-bound and memory-bound models.

Question 2
For the fastest I/O performance in TensorFlow... (check all that apply)

*A: Read TF records into your model.
Feedback: This is one of the correct answers. dataset = tf.data.TFRecordDataset(...) TF Records are set for fast, efficient, batch reads, without the overhead of having to parse the data in Python.

*B: Read in parallel threads.
Feedback: This is one of the correct answers. dataset = tf.data.TFRecordDataset(files, num_parallel_reads=40) When you're dealing with a large dataset sharded across Cloud Storage, you can speed up by reading multiple files in parallel to increase the effective throughput. You can use this feature with a single option to the TFRecordDataset constructor called num_parallel_reads.

*C: Optimize TensorFlow performance using the Profiler.
Feedback: This is one of the correct answers.

*D: Prefetch the data
Feedback: This is one of the correct answers. dataset.prefetch decouples the time data is produced from the time it is consumed. It prefetches the data into a buffer in parallel with the training step. This means that we have input data for the next training step before the current one is completed.

Question 3
What does high-performance machine learning determine?

*A: Time taken to train a model
Feedback: Correct!

B: Reliability of a model
Feedback: This answer is not correct. Review the module.

C: Deploying a model
Feedback: This answer is not correct. Review the module.

D: Training a model
Feedback: This answer is not correct. Review the module.

Question 4
Which of the following indicates that ML training is CPU bound?

A: If I/O is complex, but the model involves lots of complex/expensive computations.
Feedback: This answer is not correct. Review the module.

B: If you are running a model on powered hardware.
Feedback: This answer is not correct. Review the module.

*C: If I/O is simple, but the model involves lots of complex/expensive computations.
Feedback: Correct!

D: If you are running a model on accelerated hardware.
Feedback: This answer is not correct. Review the module.

Hybrid ML Systems

Question 1
Which of these are reasons that you may not be able to perform machine learning solely on Google Cloud? Check all that apply.

*A: You are tied to on-premises or multi-cloud infrastructure due to business reasons.
Feedback: Correct!

*B: You need to run inference on the edge.
Feedback: Correct!

C: TensorFlow is not supported on Google Cloud.
Feedback: This answer is not correct; of course Google Cloud supports TensorFlow.

Question 2
A key principle behind Kubeflow is portability so that you can:
*A: Move your model from on-premises to Google Cloud.
Feedback: Correct! Portability is at the container level, and you can move to any environment that offers Kubernetes.

B: Migrate your model from TensorFlow to PyTorch.
Feedback: This answer is not correct. Please review the module.

C: Convert your model from CUDA to XLA.
Feedback: This answer is not correct. Please review the module.

Question 3
Which of the following determines the correct property of Tensorflow Lite? Select TWO correct answers.

A: Increased code footprint
Feedback: This answer is not correct. Review the module.

*B: Quantization
Feedback: Correct!

C: Higher precision arithmetic

Feedback: This answer is not correct. Review the module.

*D: Lower precision arithmetic
Feedback: Correct.

Question 4
To copy the input data into TensorFlow, which of the following syntaxes is correct?

A: inferenceInterface.feed(floatValues, 1, inputSize, inputSize, 3);
Feedback: This answer is not correct. Review the module.

B: inferenceInterface.feed(inputName, floatValues, inputSize; inputSize);
Feedback: This answer is not correct. Review the module.

*C: inferenceInterface.feed(inputName, floatValues, 1, inputSize, inputSize, 3); Feedback: Correct!

D: inferenceInterface.feed(inputName, floatValues, 1, inputSize, 3);
Feedback: This answer is not correct. Review the module.