**1.**

Question 1

Which of the following are advantages of batch inference?

**1 / 1 point**



You can wait for data retrieval to make predictions since they are not made available in real time.

**Correct**

Excellent!



You can use complex machine learning models to improve accuracy since there is no constraint on inference time.

**Correct**

Nice job!



You achieve shorter latency of predictions.



You don’t need caching strategies so costs decrease.

**Correct**

Great work!

**2.**

Question 2

True or False: The most important metric to optimize while performing batch predictions is throughput.

**1 / 1 point**



False



True

**Correct**

Yes! Triton Inference Server allows deployment of models from any framework, from local storage, Google Cloud Platform, or AWS S3.

**3.**

Question 3

How can you save inference costs when generating recommendations on e-commerce?

**1 / 1 point**



Using hardware accelerators.



Generating them on a schedule.



Generating them every time a user logs in.

**Correct**

That's right!