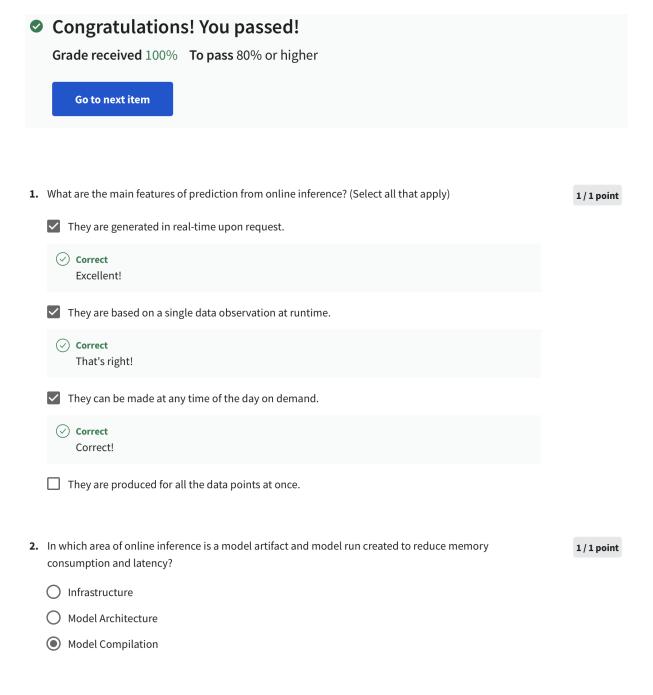
## **Online Inference**



Online Inference 1

That's right! Model Compilation are running instances of the model responsible for performing the actual inference. Thus, multiple workers can run simultaneously on Torch Serve.

**3.** True or False: Fast data caching using NoSQL databases is a cheap way of optimizing inference.

1/1 point

2

( )	True
	1140



