

 **Congratulations! You passed!**

Grade received **100%** To pass 80% or higher

[Go to next item](#)

1. What are the main features of prediction from online inference? (Select all that apply)

1 / 1 point

☒ They are generated in real-time upon request.

✓ **Correct**  
Excellent!

☒ They are based on a single data observation at runtime.

✓ **Correct**  
That's right!

☒ They can be made at any time of the day on demand.

✓ **Correct**  
Correct!

☐ They are produced for all the data points at once.

2. In which area of online inference is a model artifact and model run created to reduce memory consumption and latency?

1 / 1 point

- ☒ Model Compilation
- ☐ Model Architecture
- ☐ Infrastructure



**Correct**

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

- ☒ False
- ☐ True



**Correct**

Yes!