Scalability and Performance

10/24/2019 • 2 minutes to read •

In this article

High-performance and high-traffic websites and applications have two main factors to consider with Personalizer for scalability and performance:

- Keeping low latency when making Rank API calls
- Making sure training throughput keeps up with event input

Personalization can return a rank rapidly, with most of the call duration dedicated to communication through the REST API. Azure will autoscale the ability to respond to requests rapidly.

Low-latency scenarios

Some applications require low latencies when returning a rank. Low latencies are necessary:

- To keep the user from waiting a noticeable amount of time before displaying ranked content.
- To help a server that is experiencing extreme traffic avoid tying up scarce compute time and network connections.

Scalability and training throughput

Personalizer works by updating a model that is retrained based on messages sent asynchronously by Personalizer after Rank and Reward APIs. These messages are sent using an Azure EventHub for the application.

It is unlikely most applications will reach the maximum joining and training throughput of Personalizer. While reaching this maximum will not slow down the application, it would imply Event Hub queues are getting filled internally faster than they can be cleaned up.

How to estimate your throughput requirements

- Estimate the average number of bytes per ranking event adding the lengths of the context and action JSON documents.
- Divide 20MB/sec by this estimated average bytes.

For example, if your average payload has 500 features and each is an estimated 20 characters, then each event is approximately 10kb. With these estimates, 20,000,000 / 10,000 = 2,000 events/sec, which is about 173 million events/day.

If you are reaching these limits, please contact our support team for architecture advice.

Next steps

Create and configure Personalizer.