Есть 2 подхода:

* Сценарный – когда мы создаем сценарий (зайти на сайт, добавить товар в корзину, оплатить). Тем самым мы увидим общее время этого сценария. Но по этому подходу бывает не возможным определить на каком этапе сколько времени заняла загрузка
* Точечный (hit based) – когда мы нагружаем конкретный ендпоинт и смотрим время ответа

Виды тестов:

1. Стресс тест – мы исследуем систему на пределы (максимальное кол-во юзеров, перед тем как система откажет)
2. Нагрузочный тест – нагружаем систему тем кол-ом пользователей, которое указано в требованиях, сколько система должна обрабатывать
3. Тестирование стабильности – нагружать систему 70% нагрузки от максимальной, в течении длительного времени, что бы отследить утечки памяти
4. Восстановление – проверить как будет сервер выдерживать нагрузку, после включения, после того как он упал
5. Масштабируемость – проверить уменьшится ли время ответа и увеличится кол-во пользователей после добавления нового сервера