Desempenho

Meça o comportamento da agenda com carga controlada

Declare o cenário e as metas

Escolha uma operação com impacto para o usuário, como consultar horários. Defina a taxa de chegada, duração, dados e limites de latência e erro. No exemplo, p95 abaixo de 300 ms significa que a estimativa do percentil 95 ficou abaixo desse valor.

Compare condições iguais

Execute em ambiente autorizado de teste e registre a versão da API, a massa de dados, os recursos e o estado do cache. Compare a versão candidata com uma referência sob a mesma carga. Use uma etapa de aquecimento separada se precisar medir o serviço já aquecido.

Verifique a carga entregue

Compare a taxa observada com a planejada. Iterações descartadas podem indicar capacidade insuficiente no gerador ou lentidão no serviço. Investigue antes de interpretar a latência como representativa da taxa-alvo.

Desempenho

Teste a consulta de horários

Exemplo em k6

```
import http from "k6/http";

export const options = {
  scenarios: {
    agenda: {
      executor: "constant-arrival-rate",
      rate: 20,
      timeUnit: "1s",
      duration: "2m",
      preAllocatedVUs: 30,
    },
  },
  thresholds: {
    http_req_duration: ["p(95)<300"],
    http_req_failed: ["rate<0.01"],
    dropped_iterations: ["count==0"],
  },
};

export default function () {
  http.get(__ENV.BASE_URL + "/horarios");
}
```

Execução em homologação

Salve como horarios.js e execute k6 run -e BASE_URL=https://agenda-hml.example.com horarios.js, usando o endereço real do ambiente autorizado. O exemplo realiza uma consulta por iteração. As metas ilustrativas exigem p95 abaixo de 300 ms, menos de 1% de falhas e zero iterações descartadas.

Desempenho

Exercício resolvido

Cenário

Com 20 iterações por segundo e nenhuma descartada, a referência apresentou p95 de 210 ms e 0,2% de falhas. A candidata apresentou p95 de 340 ms e 0,4% de falhas. As duas passaram? Onde investigar?

Resolução

A referência passa os três limites. A candidata passa o limite de erros, mas excede a meta de latência em 40 ms. Correlacione o horário da medição com os traces da consulta e as métricas do banco. Se a latência estiver concentrada em uma consulta SQL, examine o plano e a cardinalidade com dados representativos.

Verificação

Repita com as mesmas condições após o ajuste. Exija os três limites e compare o resultado com a referência. Guarde o resumo do k6 e o commit usado para cada rodada.

Referências primárias

Grafana k6 — Thresholds

Grafana k6 — Iterações descartadas

Desempenho

Compare a carga planejada com a carga entregue

Leia duração e vazão em conjunto

Uma execução pode apresentar latência baixa porque entregou menos solicitações do que o planejado. Registre volume, falhas e iterações descartadas antes de comparar percentis entre versões.

Exercício adicional

O p95 caiu de 320 para 200 ms, mas houve muitas iterações descartadas. A versão melhorou?

Resolução

Primeiro identifique se o gerador ou o serviço limitou a carga. Ajuste o ensaio e repita com volume comparável. Só então compare percentis e taxa de erros.

Ao assistir ao vídeo

Identifique cenários, métricas de duração e limites do teste.

Vídeo complementar

How to load test the backend with Grafana k6 (em inglês)