Rencana Pengujian Kinerja (Performance Test Plan) - GenEdu

Anggota Kelompok:
1. Muhammad Ellbendl Satria (2023071004)
2. Gathan Ghifari Rachwiyono (2023071038)

1. Tujuan & Sasaran Pengujian

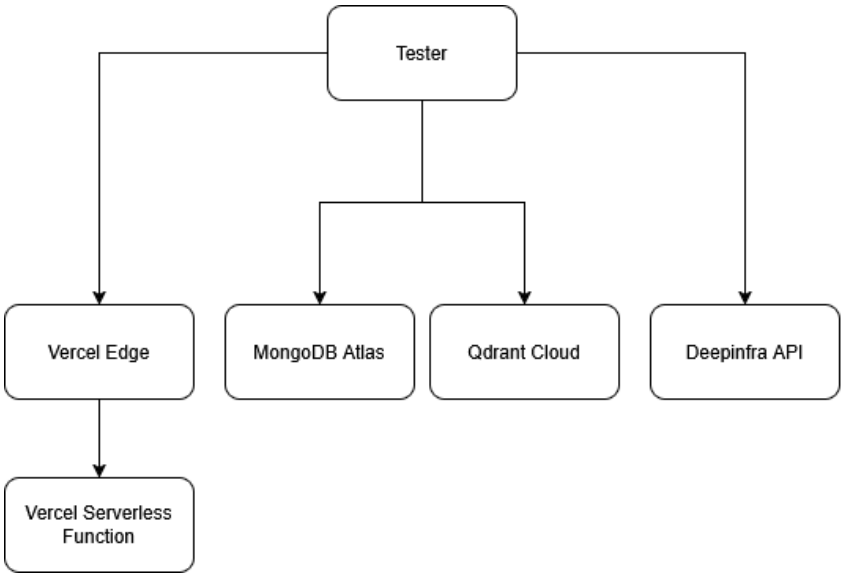
Tujuan utama dari pengujian ini adalah:

1. **Validasi Waktu Respons:** Memastikan fungsi kritis (terutama yang melibatkan AI) memenuhi target waktu respons **kurang dari 5 detik** dari perspektif pengguna.
2. **Analisis Latensi Antar Layanan:** Mengisolasi dan mengukur latensi pada setiap "lompatan" jaringan, khususnya antara Vercel dan layanan di region berbeda (Qdrant & DeepInfra di AS).
3. **Evaluasi Kinerja Serverless:** Mengukur dampak **cold start**, durasi eksekusi, dan skalabilitas Vercel Functions di bawah beban.
4. **Identifikasi Titik Batas:** Menentukan kapasitas maksimum sistem dan menemukan *bottleneck* utama dalam arsitektur.

2. Arsitektur & Lingkup Pengujian

2.1. Arsitektur Target

Pengujian akan menargetkan alur data end-to-end sebagai berikut:



2.2. Lingkup Pengujian

- **Dalam Lingkup (In-Scope):**

    - Performa *backend* pada Next.js API Routes.
    - Waktu respons untuk *Use Case* utama: Login, Unggah Materi, Buat Ringkasan, Buat Kuis, Kelola Catatan, dan Cari Info Kampus.
    - Kinerja koneksi dan *query* ke MongoDB Atlas dan Qdrant.
    - Latensi saat berinteraksi dengan API eksternal (DeepInfra).
- **Di Luar Lingkup (Out-of-Scope):**
    - Pengujian fungsionalitas (uji benar/salah).
    - Pengujian UI/UX dan rendering di sisi klien.
    - Pengujian keamanan (vulnerability scanning).

2.3. Lingkungan & Alat Uji

- **Aplikasi Target:** *Staging environment* di Vercel.
- **Database Uji:** Cluster khusus di MongoDB Atlas dan Qdrant yang diisi dengan data uji representatif.
- **Alat Generator Beban: Apache JMeter** atau **K6**, dijalankan dari server cloud di region Asia (misal: Singapura) untuk mensimulasikan lalu lintas pengguna.
- **Alat Monitoring:**
    - **Vercel:** Dashboard Analytics & Logs.
    - **MongoDB:** Dashboard Monitoring Atlas.
    - **Qdrant:** Dashboard monitoring Qdrant Cloud.
    - **Aplikasi:** Logging kustom di dalam API Routes untuk mencatat *timestamp* pada titik kritis (sebelum dan sesudah panggilan API eksternal).

3. Metrik Kinerja Utama (KPI)

Metrik	Target	Keterangan
Waktu Respons End-to-End	**< 5 detik**	Untuk proses AI (Ringkasan/Kuis).
Waktu Respons Database	**< 50 ms**	Untuk operasi CRUD sederhana ke MongoDB Atlas.
Throughput	Ditentukan	Jumlah *requests per minute* (RPM) yang dapat ditangani.
Error Rate	**< 1%**	Pada kondisi beban normal.
Vercel Cold Start Rate	Dimonitor	Persentase permintaan yang mengalami *cold start*.

4. Skenario Pengujian

Setiap skenario diawali dengan langkah **Login** untuk mendapatkan token JWT.

Jenis Tes	Skenario Pengujian	Deskripsi & Tujuan	Beban & Durasi	Metrik Utama yang Dipantau
Load Test	**Simulasi Hari**	Mengukur kinerja	100 Pengguna	Waktu Respons

	Akademik Normal	sistem di bawah beban 100 pengguna bersamaan yang melakukan berbagai aktivitas (dominan fitur AI).	Ramp-up: 10 menit Durasi: 30 menit	End-to-End, Throughput, Error Rate, Durasi Vercel Function.
Stress Test	**Identifikasi Bottleneck Antar Benua**	Meningkatkan beban secara agresif pada fungsi AI untuk menemukan titik putus akibat latensi tinggi ke layanan di AS.	Mulai 50 Pengguna Tambah 25/3 menit Durasi: Sampai gagal	Jumlah pengguna maks sebelum *error rate* > 5% atau waktu respons > 15 detik.
Spike Test	**Simulasi Cold Start Massal**	Menguji dampak lonjakan lalu lintas mendadak terhadap Vercel Functions.	Lonjakan 150 Pengguna Ramp-up: 30 detik Durasi: 10 menit	Jumlah *Cold Starts*, Waktu respons permintaan pertama vs. permintaan berikutnya.
Endurance Test	**Uji Stabilitas Koneksi & Memori**	Menjalankan beban normal dalam waktu lama untuk mendeteksi kebocoran memori atau koneksi database.	70 Pengguna Durasi: 2-4 jam	Pola penggunaan memori (Vercel) dan jumlah koneksi (MongoDB Atlas) harus stabil.

5. Analisis Bottleneck Potensial & Rekomendasi

1. **Bottleneck: Latensi Jaringan ke Qdrant & DeepInfra (AS)**
    - **Identifikasi:** Log kustom menunjukkan waktu yang sangat tinggi (>1 detik) untuk *request* ke layanan di AS.
    - **Rekomendasi Utama: Pindahkan *instance* Qdrant Cloud ke region AWS yang lebih dekat** (misal: ap-southeast-1 di Singapura). Untuk DeepInfra, implementasikan *caching* (misal: Vercel KV) untuk permintaan yang identik.
2. **Bottleneck: Cold Starts Vercel**
    - **Identifikasi:** Lonjakan waktu respons pada permintaan pertama setelah periode

tidak aktif, terutama saat *Spike Test*.
    - **Rekomendasi:** Jika menjadi masalah signifikan, pertimbangkan untuk upgrade plan Vercel yang memungkinkan konfigurasi minInstances > 0 agar fungsi tetap *warm*.

3. **Bottleneck: Rate Limit API Pihak Ketiga**

    - **Identifikasi:** Munculnya *error* 429 Too Many Requests dari DeepInfra saat *Stress Test*.
    - **Rekomendasi:** Selain *caching*, pertimbangkan untuk mengimplementasikan mekanisme antrian (*queue*) atau *retries with exponential backoff*. Upgrade paket API jika perlu.
4. **Bottleneck: Limit Koneksi MongoDB Atlas**

    - **Identifikasi:** Aplikasi mengembalikan *error* koneksi database di bawah beban tinggi.
    - **Rekomendasi:** Pastikan kode menggunakan *best practice* untuk manajemen koneksi (misal: *connection pooling*). Upgrade tier MongoDB jika volume trafik memang tinggi.

5. **Kriteria keberhasilan**

- 95% response time < 1 detik
- Error rate < 1%
- Tidak ada crash atau penurunan performa drastis