Executive Summary – Load Test Results

Objective

Assess the performance and stability of the Dummy JSON APIs under a steady, expected production-like load to verify SLA compliance and identify any early performance bottlenecks.

Response Time Trends

- Average response time: **≈ 198 ms**.
- **p95 latency** remained well below SLA (1.2 s limit).
- Response times stabilized quickly during the ramp-up phase, showing consistent performance across all endpoints.

Throughput (RPS)

- Sustained **17–18 requests per second**, close to the defined target (≥ 20 RPS).
- Throughput scaled linearly with load and maintained stability once full concurrency was reached.
- No evidence of request queuing or throttling during the test window.

Error Breakdown

Error Type	% of Errors	Cause
HTTP 200/201 (Success)	100 %	All transactions completed successfully
Assertion Warnings	< 1 %	Minor timing deviations, no functional impact
HTTP 4xx/5xx	0 %	None observed

No connection resets or socket errors were recorded.

Bottleneck Hypothesis

- Slightly lower throughput than target may result from:
- Limited backend thread or worker pool capacity, or
- JMeter client resource limits under full concurrency.
- Occasional latency spikes correspond with brief backend queuing under load.

Tuning Recommendations

- **Increase backend thread pool / connection limit** to push throughput beyond 20 RPS.
- **Optimize database query latency** for read-heavy endpoints (e.g., “Get Products”).
- **Enable persistent HTTP connections** (keep-alive) to reduce setup overhead.
- Consider **horizontal scaling** or caching once concurrency requirements grow.