ONE PAGE SUMMARY

IT & Software › API Development › Rate Limiting

Introduction to Rate Limiting

Master Rate Limiting for Efficient API Management

DEFINITION

> Rate limiting controls how often a user can access an API. It helps prevent abuse and ensures fair usage.

KEY CONCEPTS

Throttling

Throttling temporarily limits requests to prevent overwhelming the server.

Burst Capacity

Burst capacity allows a short period of high traffic before limits kick in.

Quota

A quota is a maximum number of requests allowed within a specific time frame.

HTTP Status Codes

Rate limiting often uses specific HTTP status codes like 429 to indicate limits exceeded.

Client Identification

Identifying clients through API keys or IP addresses helps enforce rate limits.

EXAMPLES

→ Limit to 100 requests per hour per user.

→ Allow 10 requests per second with a burst of 20.

→ Return 429 status code when limits are exceeded.

MEMORY TIPS

★ Think of a 'traffic cop' managing API requests.

★ Remember 'Quota' as 'Quota = Limit' to reinforce the concept.

★ Associate 'Throttling' with 'slow down' to visualize request control.

COMMON MISTAKES

✗ Confusing rate limiting with authentication.

✗ Not considering burst capacity in design.

✗ Failing to communicate limits to API users.

QUICK RECAP

Rate limiting is essential for managing API traffic and preventing abuse. Key concepts include throttling, quotas, and client identification. Always communicate limits clearly to users.