ONE PAGE SUMMARY

IT & Software › API Development › Rate Limiting

Introduction to Rate Limiting

Master Rate Limiting for Efficient API Management

DEFINITION

> Rate limiting controls the number of requests a user can make to an API in a given time frame. It helps prevent abuse and ensures fair usage among all users.

KEY CONCEPTS

Request Count

The total number of API calls made by a user within a specified period.

Time Window

The defined duration (e.g., per minute, hour) during which request counts are monitored.

Throttling

A method to slow down or limit the rate of requests from users exceeding the allowed limit.

Burst Capacity

The ability to handle a sudden increase in requests for a short time without immediate rejection.

HTTP Status Codes

Codes like 429 (Too Many Requests) indicate that a user has exceeded their rate limit.

EXAMPLES

→ A user can make 100 requests per hour.

→ An API allows 10 requests per minute with a burst capacity of 5.

→ After 429 status, users must wait before making more requests.

MEMORY TIPS

★ Think of a traffic light: green for normal requests, yellow for caution, red for stop.

★ Remember 'Limit Your Rate' to keep usage fair.

★ Use the acronym RATS: Requests, Allowed, Time, Status.

COMMON MISTAKES

✗ Confusing rate limiting with authentication methods.

✗ Not considering burst capacity when setting limits.

✗ Failing to implement proper error handling for rate limit responses.

QUICK RECAP

Rate limiting is essential for managing API traffic and ensuring fair access. By controlling request counts within specified time windows, APIs can prevent abuse and maintain performance.