Preventing system abuse: Throttling and Rate Limiting

Presenter: Rachit Sharma

/thoughtworks

What is Throttling ?

I see it as defense mechanism where the rate of requests/data sent is at the acceptable rate as per the target systems capacity.

Why throttle or Rate Limit ?

Slow, Reject, Ignore

Prevent System from getting down, 429 better than 503, Acceptable rate of consumption , acting as a buffer.

Controls Virality, Prevent accidental surges

It's better to ask users to visit post sometime rather than system being down

Could possibly ignore the DDOS by ignoring the requests as per certain rules, like sending the attacker 200 for multiple requests via same IP.

Control Infra cost, might be using a costly AI/ML api.

Scenarios

🚩	A user has gone rogue and is madly hitting our api
🌐	Virality
🚩	Control Infra/consumption cost internally
🌐	Overwhelmed Database

A rogue user

1. Interfaces rightly with the api/public facing endpoint
2. Have a bucket of tokens associated with the user / auth token
3. For each bucket have specified number of tokens
4. For each request deduct the tokens
5. Refill bucket at periodic intervals
6. If before the refill user exhausts all tokens then return 429

Overwhelmed database

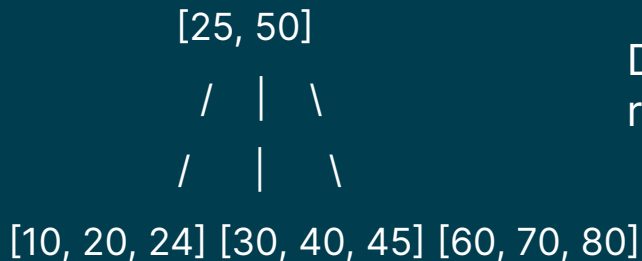Having public facing api that deletes something from database

Frequent hits to database during write heavy applications

Hard deletes vs Soft Deletes

1. Locking
2. Cascade deletes and referential integrity overhead
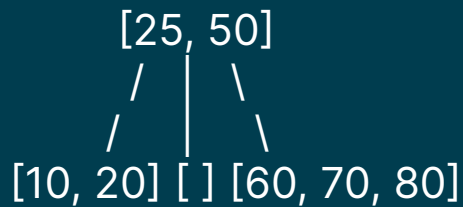3. Index Management
   a. update, rebalance, fragmentation
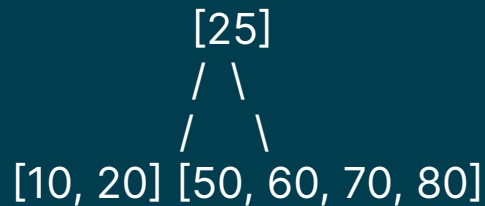4. WAL

Index Issues

Initial B+ tree

```
        [25, 50]
        /   |   \
       /    |    \
[10, 20, 24] [30, 40, 45] [60, 70, 80]
```

DELETE action to remove 30,40,45

Post deletion

```
        [25, 50]
       /   |   \
      /    |    \
[10, 20] [ ] [60, 70, 80]
```

Rebalancing

```
        [25]
        /  \
       /    \
[10, 20] [50, 60, 70, 80]
```

Mentions

1. WAL : Brewing Implementation for Write-Ahead Logs | by Rachit Sharma | Jun, 2024 | Medium

Thank You!