Rest - Fast API For ML in Production

Furkan Ataç

Outline

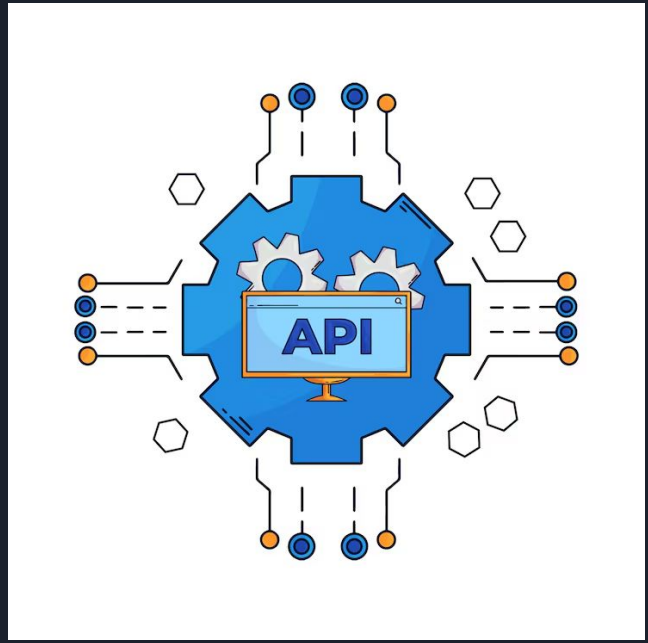
- What is an API
- Why API?
- HTTP API
- REST API
- FAST API
- MLOps Fit
- Conclusion

What is an API?

- A Contract that Softwares can talk to
- Request then response
- Interface, abstraction, schema
- May be in various formats
- Client - API - Database/Service/Model

Why Do We Need APIs?

- Reuse or scale
- Security and Governance
- Productable
- Separate boundaries



HTTP API

- GET - POST - PUT - DELETE (CRUD operations we see)
- Routes are simple: /resource/<id>
- Status codes: 200 (successful) and various errors
- JSON supported

Rest API

- Resource centric
- Stateless - therefore easy to scale
- Uniform interface and caching
- Simple

Fast API

- Speedy and Async for high throughput
- Clear errors, testable endpoints
- Define request/response once, fast api handles scalability
- Auto-docs with OpenAI
- Clean wiring to databases
- Strict schema -> fewer data-drift surprises (beneficial for MLOps)

Why This Matters for MLOps

- Versioning well
- Observability
- Auto-scaling
- Security
- Reproducibility

Conclusion

- API = Contract
- HTTP/REST - common language
- FastAPI - Python - Production Service
- Production Ready
- Scalable
- Github URL