

KEA

API Design

Product description

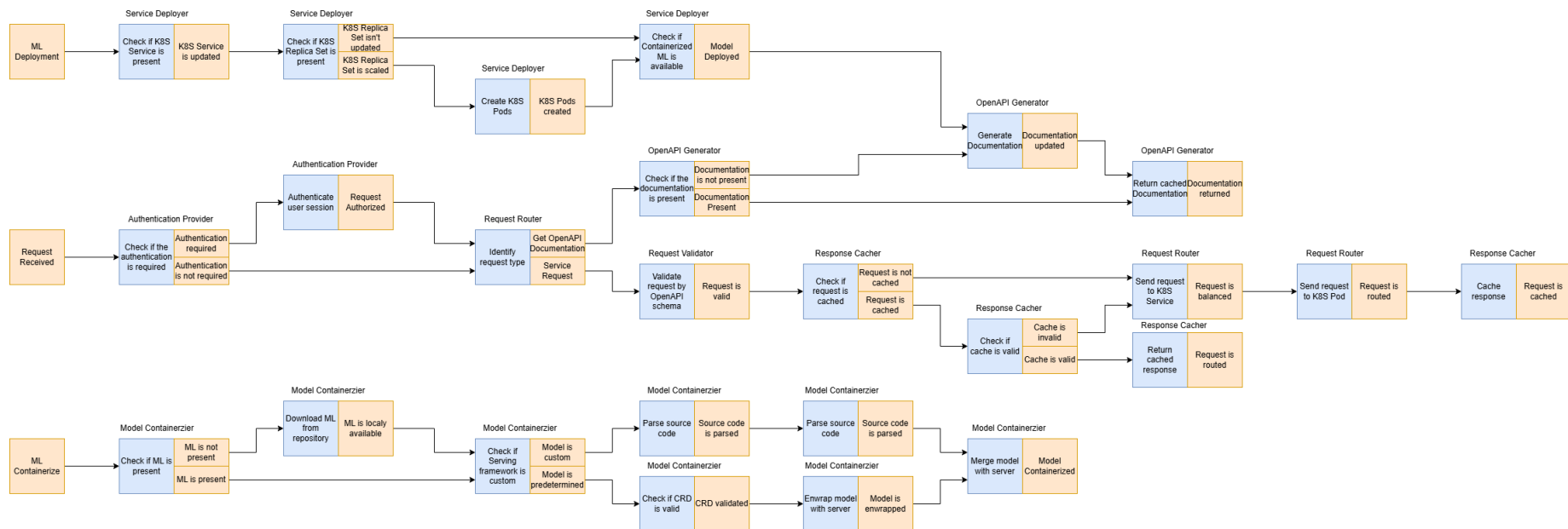
The product is a platform for deploying, managing, and scaling machine learning models in production. It offers a secure, flexible environment for automating ML tasks like model versioning, routing, and monitoring. With Kubernetes integration and containerization support, it's designed for developers, ML engineers, and enterprises needing scalable, reliable ML infrastructure.

Team K8C: Tsurkan Daniel; Dandamaev Gadji; Tsaturyan Konstantin; Smolkin Mikhail

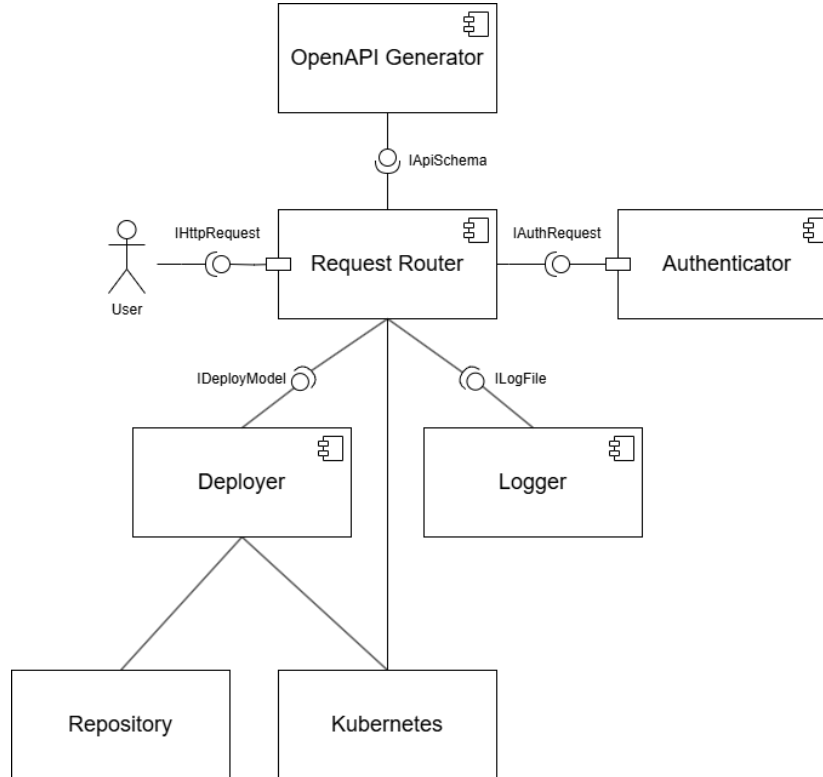
Project repo: <https://github.com/fanglores/Advanced-Software-Design>

This report: [https://github.com/fanglores/Advanced-Software-Design
/blob/master/Practice%20Tasks/Module2/Task_10/Task_10.pdf](https://github.com/fanglores/Advanced-Software-Design/blob/master/Practice%20Tasks/Module2/Task_10/Task_10.pdf)

Event flow



Service diagram



Open API

Paths:

- /router/loadbalancer
- /router/loadbalancer/{serviceId}
- /services
- /services/{serviceId}
- /services/{serviceId}/predict
- /auth/login
- /openapi/{serviceId}
- /logs/{serviceId}
- /router/validate
- /router/cache

Schemas:

- LoadBalancerConfig
- LoadBalancerResponse
- LoadBalancerInfo
- LoadBalancerDetails
- DeployRequest
- DeployResponse
- SenriceInfo
- ServiceDetails
- UpdateRequest
- UpdateResponse
- PredictRequest
- PredictResponse
- LoginRequest
- LoginResponse
- Logs
- Validation Request

API usage **ServiceDeployer**

Use Case: Deploy Service

Scenario:

User sends request to
deploy ML Model

ML Wrappers creates
docker container

Docker container is
deployed via service into
K8s

ServiceDeployer Deploys and manages ML services. ^

POST

/services Deploy an ML service



GET

/services List all deployed services



GET

/services/{serviceId} Get service details



PUT

/services/{serviceId} Update an ML service



DELETE

/services/{serviceId} Delete an ML service



POST

/services/{serviceId}/predict Make a prediction



API usage Authenticator

Use Case: Authenticate

Scenario:

User sends request

Request is sent to
authentication

Request is checked for
SSO authentication
possibility

Authenticator Handles authentication and RBAC validation.



POST

/auth/login Perform SSO authentication



API usage OpenAPIGenerator

Use Case: Generate
OpenAPI Schema

Scenario:

User sends request to get
schema

OpenApi Generator checks
if schema is present, if not
– creates it, returns actual
schema

OpenAPIGenerator Generates OpenAPI schemas for ML services. ^

GET

/openapi/{serviceId} Get the OpenAPI schema of a deployed service

✓

API usage **Logger**

Use Case: Logging

Scenario:

User send request to fetch
logs

Logger retrieves logs for a
service

Logger Collects and provides logs for deployed services. ^

GET

/logs/{serviceId} Fetch logs for a deployed service



API usage RequestRouter

Use Case: Forward Request

Scenario:

User sends request to a service

Request is being validated by OpenAPI schema

Request is being forwarded to a specific K8s service

RequestRouter Routes and validates API requests, with caching support. ^

POST

/router/loadbalancer Configure load balancer for a deployed ML service



GET

/router/loadbalancer List all load balancer configurations



GET

/router/loadbalancer/{serviceId}
Get load balancer configuration for a specific ML service



DELETE

/router/loadbalancer/{serviceId} Remove load balancer configuration



POST

/router/validate Validate an API request against OpenAPI schema



GET

/router/cache Fetch cached API response



Solution stack (prepare)

Implementation

- API definition: OpenAPI
- Connection server for API: python gunicorn
- App framework: python FastAPI
- Serialization/state format: json

Testing tools pytest

Operations

- App initializer: cookiecutter
- Code build: makefile
- CI/CD pipeline: github ci/cd
- Delivery method: docker
- Logging & monitoring: ELK