IntelliDoc: Cloud-native AI-assisted Multi-person Real-time Collaboration Document System Based On K8s

Boxuan Hu, Xiliang Xian, Xinyuan Xia, Shuyang Zhou

DATA SCHOOL OF COMPUTING

NUS National University of Singapore

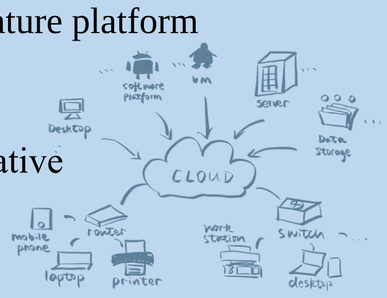
Project Scope and Motivation

Scope:

- Generative AI has been proven effective in general field, but **make mistakes** in particular working scenario
- **Effective multi-person collaboration** is in great need but lacks mature platform

Objective:

- Leveraging generative AI's power in collaborative scenarios to altimately accomplish **crowd sourcing**



- Building **a General cloud-native finetune AI architecture** for different usage scenarios
- Exploiting kubernetes' features as scalability, self-healing, batch execution, automatic bin rollouts, storage orchestration, etc.

Cloud-native Architecture

Platforms:

- Images: Docker
- Manager: Kubernetes (minikube locally)
- Cloud: aws: eks,ec2…

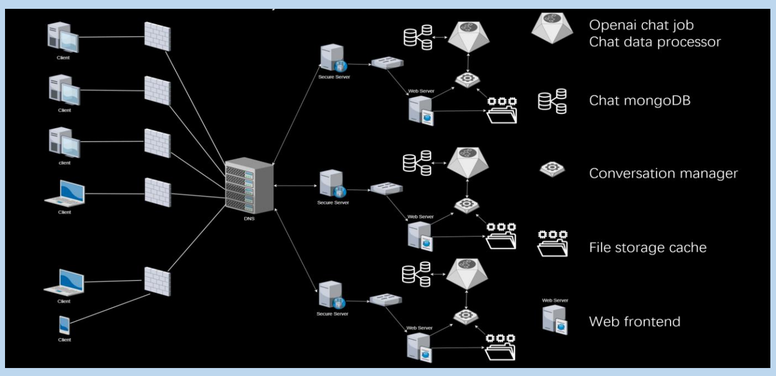
Advantages:

- High coverage and scalability
- High Availability
- Efficient Data Processing

Pipeline:

- multiple users visit the same web page, with all other interfaces behind the scene on the cloud

Client-Service Mindmap



Client:

- Multiple users send requests through frontend webpage by http proxy

Service:

- Different k8s objects handle the request
- MongoDB database sends all correction data for finetune on a daily basis
- Cloud Drive keeps all document files

Methodology and Pod-wise Mindmap

Cloud Drive Server Pod:

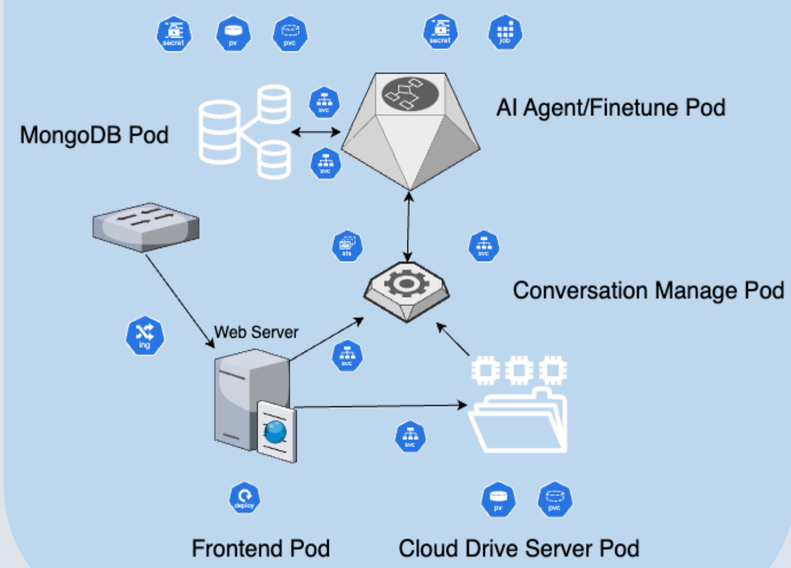
- Persistant storage for all the codes, files, etc.
- implemented by PV

Frontend Pod:

- interaction interface for operations like creating a new file/folder, editing files, generating AI response
- the basic collaboration interface of all users

Conversation Manager Pod:

- Cleanses, formats, and performs preliminary analysis on user input documents and dialogue
- Supports parallel processing, increasing system throughput
- Dynamic scaling based on load



OpenAI Interaction Pod:

- Interacts with the OpenAI API to handle AI-assisted requests.
- Safely stores the API with secrets

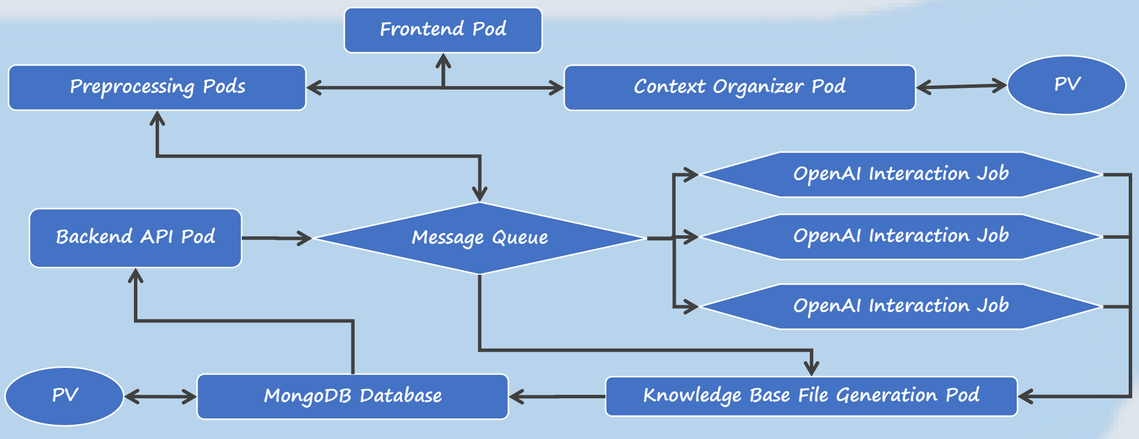
MongoDB Document Database:

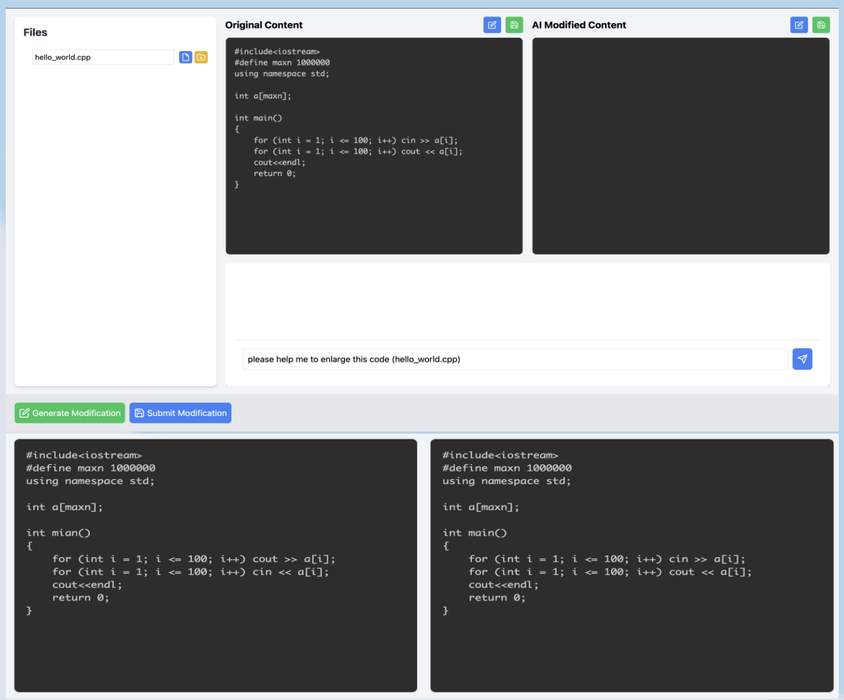
- Persistent storage for history conversations and file edition

Results and Discussion

Features:

- Basic features as folders and files creating, real-time rendering, editing, Cloud Drive
- Tailored AI assistant for your whole work group
- Easy editing AI-generated coding, with just one click





Highlights:

- **High Scalability** with kubernetes' horizontal scaling
- **Robust Processing Capabilities** with high-throughput real-time data stream processing
- **High Availability and Fault Tolerance** with key pods scaled
- **Optimized Performance** with paralized data preprocess, caching and Jobs implemented
- **Maintenance Friendly** with our micro-service design

Future work:

- Building a General cloud-native finetune AI architecture
- Using better ai-models

Visit our github repository for tutorial on implementation and explore fancy functions!