

Document Clustering System with Docker

Technical Mathematics for Big Data

Oyedotun Oluwasegun Michael (#123168)

Silvia Mastracci (#123177)

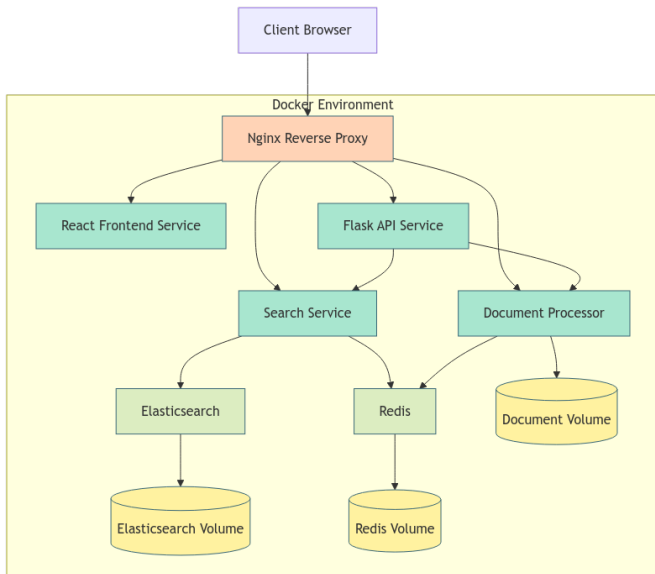
Oleksandr Solovei (#126784)

January 16, 2025

Project Overview

- Document Clustering System built with microservices
- Containerized solution using Docker technology
- Key technical components:
 - Frontend: React-based UI
 - Backend: Flask microservices
 - Search: Elasticsearch engine
 - Processing: Document analysis system

System Architecture



Docker Compose Configuration

```
version: '3.8'
services:
  nginx:
    image: nginx:alpine
    ports:
      - "4321:80"
  frontend:
    build: ./frontend
    expose:
      - "3000"
  api:
    build: ./api
    expose:
      - "8000"
  document-processor: (...)
  search: (...)
```

Running the System

Running commands:

Build and start

```
docker-compose up --build
```

Stop services

```
docker-compose down
```

View logs

```
docker-compose logs -f
```

Rebuild specific service

```
docker-compose build service-name
```

- **Consistent Development Environment**

- Same environment for all team members
- "Works on my machine" problem eliminated
- Quick onboarding of new developers

- **Isolated Dependencies**

- Each service has its own container
- No conflicts between different versions
- Easy technology stack updates

- **Rapid Development Cycle**

- Fast container startup
- Quick iteration and testing
- Easy rollback capabilities

• Resource Efficiency

- Lightweight container architecture
- Optimal resource utilization
- Lower infrastructure costs

• Scalability

- Easy horizontal scaling
- Load balancing support
- Dynamic resource allocation

• Maintenance

- Simple updates and patches
- Minimal downtime
- Easy backup and restore

- **Container Isolation**

- Separate process spaces
- Independent network interfaces
- Isolated file systems

- **Security Features**

- Resource limitations
- Capability restrictions
- Network security policies

- **Vulnerability Management**

- Container image scanning
- Regular security updates
- Immutable infrastructure

Project Implementation

- **Multi-stage builds**

- Optimized image sizes
- Reduced attack surface

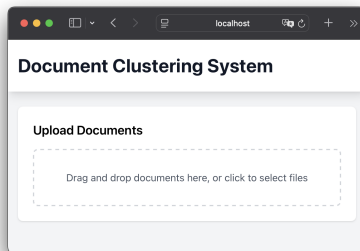
- **Docker Compose**

- Service orchestration
- Environment configuration
- Network management

- **Volume Management**

- Persistent data storage
- Efficient data sharing

| <input type="checkbox"/> | Name | Tag | Image ID |
|--------------------------|---|--------|--------------|
| <input type="checkbox"/> | redis | alpine | 1bf97f21f01b |
| <input type="checkbox"/> | docker.elastic.co/elasticsearch/elasticsearch | 7.17.9 | 59b37f77bd8b |
| <input type="checkbox"/> | nginx | alpine | 814a8e88df97 |
| <input type="checkbox"/> | docker-project-search | latest | 4f5fbb64e1de |
| <input type="checkbox"/> | docker-project-document-processor | latest | d8329a9c19bc |
| <input type="checkbox"/> | docker-project-api | latest | 1806fbbc4c1a |
| <input type="checkbox"/> | docker-project-frontend | latest | a966962ace67 |



- **Single Port Access**

- All services through one port
- Nginx reverse proxy
- Simplified deployment

- **Monitoring**

- Health checks
- Service discovery
- Automated recovery

- **Data Management**

- Elasticsearch integration
- Redis caching
- Persistent storage

Thank You

Document Clustering System
Docker-based Microservices Architecture

Questions?