Development Infrastructure

McMillan Project Documentation

September 1, 2025

Overview

The development infrastructure is designed to support a modular, containerized application that integrates backend services, AI-based document processing, a relational database, and frontend applications. It leverages **AWS Cloud resources** for scalability, cost efficiency, and reliability.

1 Core Components

1.1 Backend (FastAPI)

- Hosted in Docker container (`app` service)
- Handles REST API requests, authentication, and communication with database
- Uses environment variables and `.env` file for configuration
- Logs are persisted to host machine (`./app/logs:/app/logs`)

1.2 InvoiceAI Service

- Custom container (`consulttechencraft/invoiceai`)
- Communicates with **OpenAI API** for AI-driven document analysis
- Handles OCR/text processing before results are sent to Postgres
- Runs independently but shares the same **Docker network** for inter-service communication

1.3 Database (Postgres)

- Dockerized Postgres 15
- Persistent storage via Docker volume `db_data`
- Exposed on port `5432` for backend connectivity
- Health checks ensure DB is running before dependent services start

1.4 pgAdmin

- UI management tool for Postgres
- Runs on Docker container (`pgadmin` service)
- Accessible via port `5050`
- Uses admin credentials defined in `.env`

1.5 🌐 Frontend Applications

- Two static web applications (e.g., React/Angular/Vue builds)
- Hosted in **AWS S3 buckets** with static website hosting enabled
- Delivered via **S3 public URLs or CloudFront (optional CDN)**

2 AWS Infrastructure

2.1 EC2 Instance

- **t3.medium** (2 vCPUs, 4 GiB RAM)
- Runs Docker + Docker Compose
- Containers: FastAPI backend, InvoiceAI, Postgres, pgAdmin
- Cost optimized: Runs ~10 hrs/day (~$12.5/month)

2.2 EBS (Elastic Block Store)

- 20 GiB gp3 volume for EC2 instance storage
- Persists database and application logs

2.3 S3 Buckets

- **10 GiB S3 storage** for document data
- **2 S3 buckets** for frontend apps hosting
- Cost-effective and highly available

3 Networking & Security

- **VPC** with private Docker bridge network (`mcmillan-net`) for service communication
- **Public access** only for:
  - Backend API (port 8000)
  - InvoiceAI service (port 8001)
  - pgAdmin (port 5050)
- Security Groups restrict database exposure to internal network only
- **IAM roles & policies** manage S3 access

4 Monitoring & Logging

- **CloudWatch** for EC2 instance metrics (CPU, memory, disk usage)
- **FastAPI + InvoiceAI logs** persisted in `./app/logs`
- **Postgres logs** available via container logs

Component	Usage	Est. Cost
EC2 (t3.medium, 10 hrs/day)	Compute	$12.48
EBS (20 GiB gp3)	Storage	$1.60
S3 (10 GiB backend data)	Storage	$0.23
S3 (2 frontend apps + traffic)	Hosting & transfer	$0.85
Total		**$15.2**

Table 1: Monthly cost breakdown for development infrastructure

5 $ Cost Summary (Monthly)

6 ↑ Scalability Path

Short Term

t3.medium is sufficient for current requirements.

If Traffic Increases

- Upgrade to **t3.large (8 GiB RAM)**
- Offload Postgres to **Amazon RDS** for managed performance
- Add **CloudFront** for frontend apps to improve latency

If AI Workload Grows

Switch to containerized GPU inference (ECS/EKS) or continue relying on OpenAI API.

Summary

This setup ensures a **balanced development environment**:

- Low cost (∼$15/month)
- Containers provide isolation & easy deployment
- AWS services (EC2, S3, EBS) provide scalability & persistence