Document Management System

Cloud-based DMS with AI-powered Processing

Willian Pinho | Technical Assessment

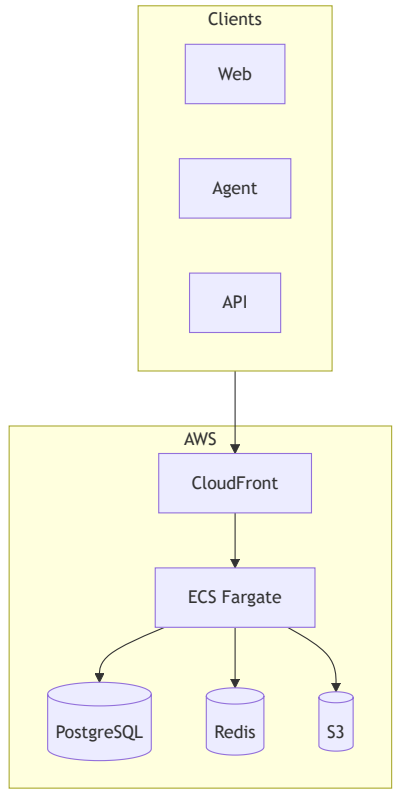
High-Level Architecture

Cloud: AWS

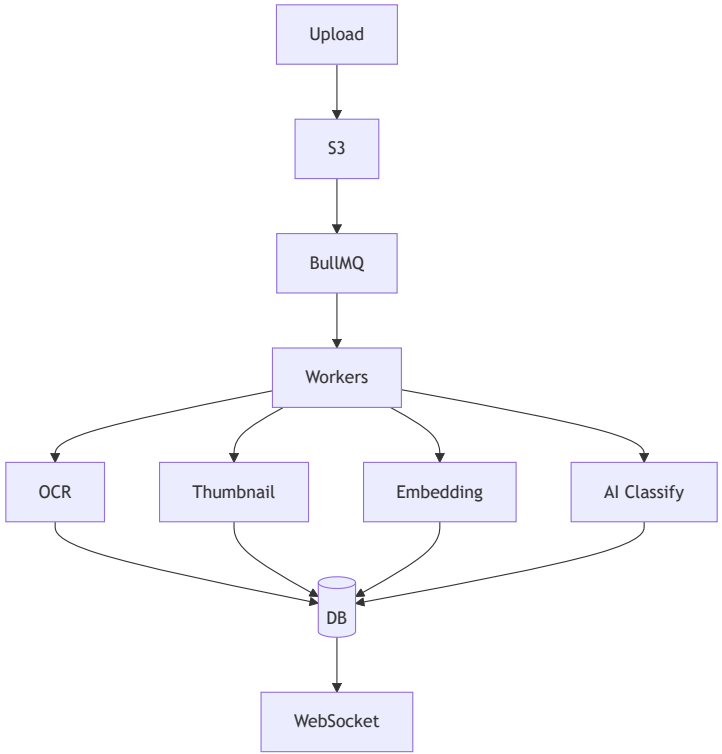
- **ECS Fargate** - Auto-scaling containers
- **PostgreSQL + pgvector** - DB + Vector search
- **S3 + CloudFront** - Storage & CDN
- **BullMQ + Redis** - Job processing

Clients

- Web App (Next.js 15)
- Desktop Agent (Electron)
- REST API



Document Processing Pipeline



Processing Jobs

- **OCR** - AWS Textract extracts text
- **Thumbnail** - Sharp generates previews
- **Embedding** - OpenAI 1536-dim vectors
- **AI Classify** - GPT-4 categorization

Features

- Presigned URLs (direct S3 upload)
- Real-time status via WebSocket
- Auto-retry with dead-letter queue
- Horizontal scaling of workers

AI Features: Search & Classification

Semantic Search

1. Document → Text extracted (Textract)
2. Text → Embedding (OpenAI)
3. Embedding → pgvector (1536 dims)
4. Query → Cosine similarity
5. Ranked results returned

Search "employee guidelines" → Finds "Company Handbook.pdf" (81% match)

AI Classification (GPT-4)

```
{
  "category": "Invoice",
  "confidence": 0.99,
  "tags": ["payment", "billing"],
  "summary": "Invoice for...",
  "language": "en"
}
```

Categories: Invoice, Contract, Report, Manual, Memo, Legal

95-99% confidence on classification

Technology Stack

Layer	Technology	Rationale
Frontend	Next.js 15 + React 19	SSR, App Router, Performance
Backend	NestJS 11	TypeScript, Modular, Enterprise
Database	PostgreSQL + pgvector	ACID + Vector Search in one DB
Queue	BullMQ + Redis	Reliable, priorities, scheduling
Storage	S3 + CloudFront	Scalable, CDN, cost-effective
AI	OpenAI + AWS Textract	GPT-4 + Enterprise OCR
IaC	AWS CDK v2	Type-safe TypeScript

Security & Observability

Authentication

- **Web App**: NextAuth.js with OAuth
  - Google & Microsoft SSO
  - JWT in HTTP-only cookies
- **API**: Bearer tokens (15min / 7day)
- **Upload Agent**: API Key + HMAC
- **RBAC**: Viewer → Editor → Admin → Owner

Observability

- **Logs**: CloudWatch structured logs
- **Metrics**: Custom CloudWatch metrics
- **Tracing**: AWS X-Ray distributed tracing
- **Alerts**: SLO-based alerting
  - P99 latency < 500ms
  - Availability > 99.9%
  - Error rate < 0.1%

Scalability & Cost

Auto-Scaling

- **ECS Fargate** scales on CPU/memory
- **Worker Spot instances** - 70% savings
- **PostgreSQL read replicas**
- **Redis cluster mode**

Development Process

- **AI-Assisted**: Claude Code + Copilot
- **Monorepo**: Turborepo for builds
- **Type Safety**: End-to-end TypeScript
- **Testing**: Vitest + Playwright

Monthly Cost (1000 users)

$500-800

Service	Cost
ECS Fargate	~$200
RDS PostgreSQL	~$150
ElastiCache Redis	~$50
S3 + CloudFront	~$50
AI APIs	~$100-200

70% savings with Spot instances

Summary

What I Built

- ✅ Production-ready DMS
- ✅ AI-powered semantic search
- ✅ Automatic document classification
- ✅ Real-time collaboration
- ✅ Complete IaC (AWS CDK)

9 Aspects Covered

- 🛠️ AI-assisted development
- ☁️ AWS cloud platform
- ⚡ ECS Fargate compute
- 🔄 GitHub Actions CI/CD
- 📊 CloudWatch observability
- 💻 Next.js + NestJS stack
- 📄 BullMQ file processing
- 🔐 OAuth + RBAC auth
- 🐘 PostgreSQL + pgvector

Ready for deep-dive discussion!

Thank You!

Willian Pinho

Technical Assessment - CultureEngine