Oil & Gas Data Management Platform - System Design

Executive Summary

A scalable web application for processing handwritten tabular data from oil fields using OCR technology, with real-time dashboard capabilities and collaborative data editing features.

Technology Stack

Frontend

- **React 18** with TypeScript for type safety
- **Next.js 14** for SSR, routing, and performance optimization
- **Tailwind CSS** for responsive design
- **Recharts** for data visualization
- **React Query** for efficient data fetching and caching
- **Socket.io-client** for real-time updates

Backend

- **Node.js** with **Express.js** and TypeScript
- **Prisma ORM** for database management
- **Socket.io** for real-time communication
- **Bull Queue** with Redis for background job processing
- **JWT** for authentication
- **Multer** for file upload handling

Database & Storage

- **PostgreSQL** for structured data storage
- **Redis** for caching and session management
- **AWS S3** for image storage
- **CloudFront** for CDN

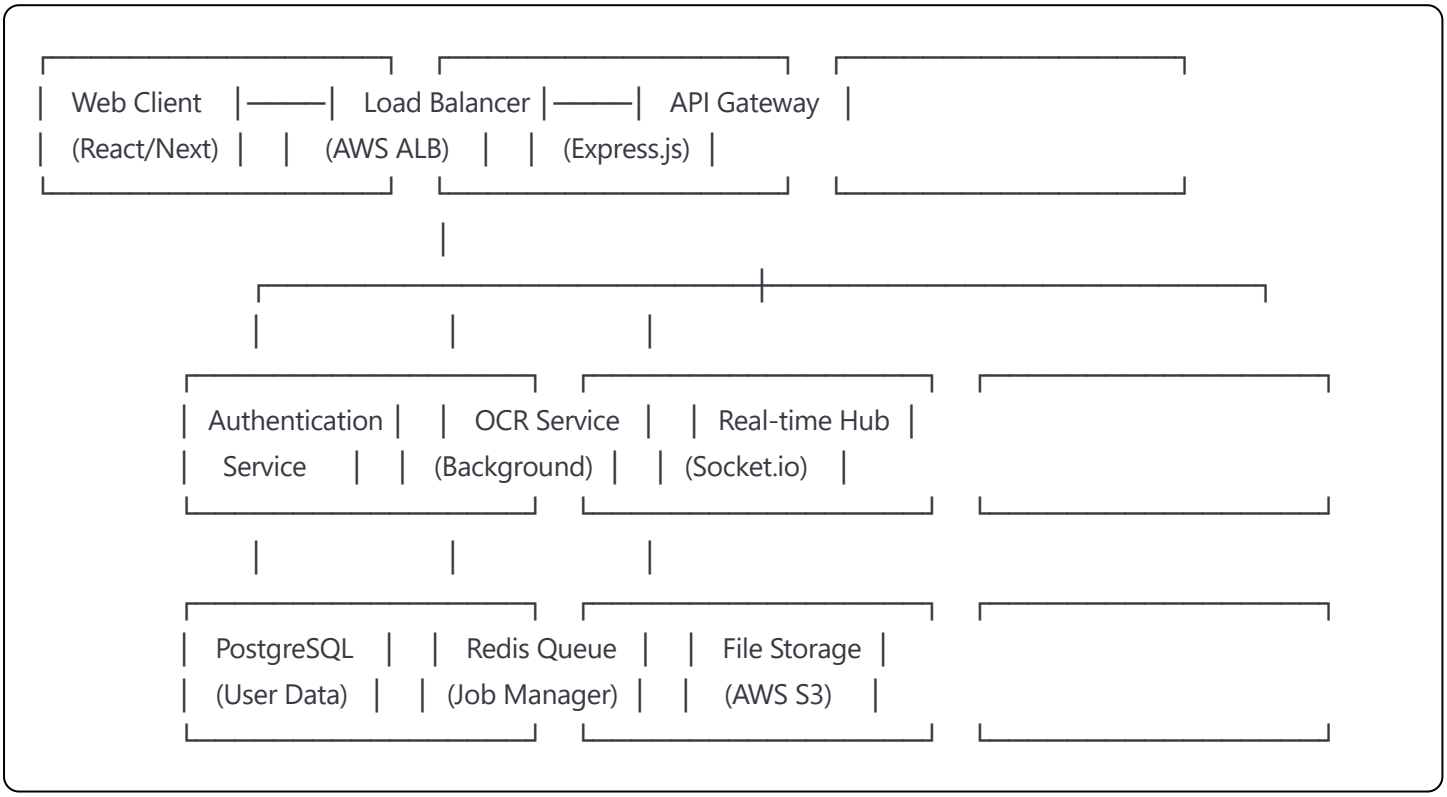
OCR & Image Processing

- **Google Cloud Vision API** (primary OCR)
- **Tesseract.js** (fallback OCR)
- **Sharp** for image preprocessing
- **OpenCV** for advanced image enhancement

Infrastructure

- **Docker** for containerization
- **AWS ECS** for container orchestration
- **AWS RDS** for managed PostgreSQL
- **AWS ElastiCache** for managed Redis
- **GitHub Actions** for CI/CD

System Architecture



Core Components

1. Authentication & User Management

- JWT-based authentication with refresh tokens
- Role-based access control (Admin, Field Engineer, Viewer)
- Multi-tenant architecture for different oil companies

2. Image Upload & Processing Pipeline

- Multi-file drag-and-drop interface
- Image validation and preprocessing
- Asynchronous OCR processing with job queues
- Progress tracking and error handling

3. OCR Engine

- Primary: Google Cloud Vision API for high accuracy
- Fallback: Tesseract for cost optimization
- Custom table detection algorithms
- Post-processing for data validation

4. Data Management

- Structured storage with field relationships
- Version control for data edits
- Audit trails for compliance
- Bulk import/export capabilities

5. Real-time Dashboard

- Live data updates via WebSockets
- Customizable chart configurations
- Drill-down capabilities
- Export functionality (PDF, Excel)

Database Schema

Core Tables

```sql
```

```sql
-- Users table
CREATE TABLE users (
  id UUID PRIMARY KEY DEFAULT gen_random_uuid(),
  email VARCHAR(255) UNIQUE NOT NULL,
  password_hash VARCHAR(255) NOT NULL,
  role VARCHAR(50) NOT NULL DEFAULT 'viewer',
  company_id UUID REFERENCES companies(id),
  created_at TIMESTAMP DEFAULT NOW()
);

-- Oil fields table
CREATE TABLE oil_fields (
  id UUID PRIMARY KEY DEFAULT gen_random_uuid(),
  name VARCHAR(255) NOT NULL,
  location VARCHAR(255),
  company_id UUID REFERENCES companies(id),
  created_at TIMESTAMP DEFAULT NOW()
);

-- Daily reports table
CREATE TABLE daily_reports (
  id UUID PRIMARY KEY DEFAULT gen_random_uuid(),
  oil_field_id UUID REFERENCES oil_fields(id),
  report_date DATE NOT NULL,
  status VARCHAR(50) DEFAULT 'processing',
  uploaded_by UUID REFERENCES users(id),
  created_at TIMESTAMP DEFAULT NOW()
);

-- Extracted data table
CREATE TABLE field_data (
  id UUID PRIMARY KEY DEFAULT gen_random_uuid(),
  report_id UUID REFERENCES daily_reports(id),
  parameter_name VARCHAR(255) NOT NULL,
  parameter_value DECIMAL(15,4),
  unit VARCHAR(50),
  confidence_score DECIMAL(3,2),
  is_verified BOOLEAN DEFAULT FALSE,
  verified_by UUID REFERENCES users(id),
  created_at TIMESTAMP DEFAULT NOW()
);

-- Image metadata table
CREATE TABLE uploaded_images (
  id UUID PRIMARY KEY DEFAULT gen_random_uuid(),
  report_id UUID REFERENCES daily_reports(id),
```

```
  file_path VARCHAR(500) NOT NULL,
  original_filename VARCHAR(255),
  file_size INTEGER,
  ocr_status VARCHAR(50) DEFAULT 'pending',
  created_at TIMESTAMP DEFAULT NOW()
);
```

Key Challenges & Solutions

1. OCR Accuracy on Handwritten Data

Challenge: Handwritten text recognition is inherently less accurate than printed text.

Solutions:

- Image preprocessing (noise reduction, contrast enhancement, deskewing)
- Multiple OCR engines with confidence scoring
- Custom training models for oil industry terminology
- Human-in-the-loop validation workflow

2. Real-time Data Synchronization

Challenge: Multiple users editing data simultaneously.

Solutions:

- Optimistic locking with conflict resolution
- Operational transform for real-time collaboration
- Event sourcing for audit trails
- WebSocket-based live updates

3. Scalability & Performance

Challenge: Processing large volumes of images and data.

Solutions:

- Horizontal scaling with container orchestration
- Background job processing with queues
- Database sharding by company/region
- CDN for image delivery
- Caching strategies at multiple levels

4. Data Quality & Validation

Challenge: Ensuring extracted data accuracy and consistency.

Solutions:

- Multi-stage validation pipeline
- Statistical anomaly detection
- Cross-field validation rules
- Manual review workflows
- Data quality scoring

Security Considerations

Data Protection

- End-to-end encryption for sensitive data
- Field-level encryption for critical parameters
- Secure image storage with signed URLs
- Regular security audits and penetration testing

Access Control

- Multi-factor authentication
- Role-based permissions with fine-grained controls
- API rate limiting and DDoS protection
- Session management with secure tokens

Compliance

- SOC 2 Type II compliance
- GDPR compliance for EU operations
- Industry-specific regulations (if applicable)
- Regular backup and disaster recovery testing

Performance Optimization

Frontend

- Code splitting and lazy loading
- Image optimization and lazy loading
- Service worker for offline capabilities
- Progressive Web App features

Backend

- Database query optimization with proper indexing
- Connection pooling and query caching
- Horizontal scaling with load balancing
- Background job processing for heavy operations

Monitoring & Observability

- Application performance monitoring (APM)
- Real-time error tracking
- Custom metrics for OCR accuracy
- User experience monitoring

Deployment Strategy

Development Pipeline

```yaml
Development → Testing → Staging → Production
    ↓          ↓       ↓         ↓
  Local      Unit Tests  E2E Tests  Blue-Green
  Docker     Integration Load Tests Deployment
  Compose    Tests       Security
                         Scans
```

Infrastructure as Code

- Terraform for AWS resource provisioning
- Helm charts for Kubernetes deployments
- Environment-specific configurations
- Automated backup and monitoring setup

Cost Optimization

OCR Processing

- Intelligent routing between premium and free OCR services
- Batch processing during off-peak hours
- Image compression without quality loss
- Caching OCR results for similar images

Infrastructure

- Auto-scaling based on demand
- Spot instances for background processing
- Reserved instances for baseline capacity
- Regular cost analysis and optimization

Future Enhancements

Advanced Analytics

- Machine learning for predictive maintenance
- Anomaly detection in production data
- Automated report generation
- Integration with IoT sensors

Mobile Applications

- Native mobile apps for field data collection
- Offline data entry capabilities
- Voice-to-text for hands-free operation
- GPS integration for location tracking

Integration Capabilities

- REST and GraphQL APIs
- Webhook support for external systems
- Third-party integrations (SAP, Oracle)
- Data export to analytics platforms