🤖 Document Intelligence Pipeline

AI-Powered Legal Analytics POC with Real Results

📋 The Problem

Legal firms face challenges with:

- ✓ Manual processing of diverse documents
- ✓ Time-consuming data extraction
- ✓ Inconsistent data quality
- ✓ Limited analytics capabilities

💡 Our Solution

Automated AI pipeline that:

- ✓ Ingests PDF documents automatically
- ✓ Classifies by type (95% accuracy)
- ✓ Extracts structured fields with LLM
- ✓ Validates and stores data

🛠️ Technical Stack (POC)

📄 **pdfplumber**
Text Extraction

🤖 **Qwen 2.5 (7B)**
Local LLM

✅ **Pydantic**
Validation

🐼 **Pandas**
Processing

🎯 POC Results - Real Data Processed

Document	Type	Vendor/Source	Amount	Confidence
uber.pdf	**Invoice**	Uber Technologies	$64.46 CAD	95%
wework.pdf	**Invoice**	WeWork Canada LP	$36.75 CAD	95%
cargo.pdf	**Invoice**	Cargo Collective	$99.00 USD	95%

3
Documents

95%
Confidence

100%
Success

$0
API Costs

Detailed Execution Pipeline

5-Stage Architecture with Real Sample Output

5-Stage Pipeline

STAGE 1: INGESTION

Input: case_dataset.pdf (3 pages) → Output: Text + metadata

STAGE 2: CLASSIFICATION

Qwen 2.5 analyzes → All 3 classified as "invoice" (95%)

STAGE 3: EXTRACTION

Invoice prompts → Client, amount, dates, vendors extracted

STAGE 4: VALIDATION

Pydantic validates → Type-safe Invoice objects

STAGE 5: STORAGE

Save to JSON + CSV → data/output/*.json + *.csv

Sample Output: WeWork Invoice

Invoice #:	PXC7PUAWY2HY-1
Date:	June 17, 2025
Client:	Pentcho Tchomakov
Vendor:	WeWork Canada LP
Amount:	$36.75 CAD
Status:	PAID IN FULL

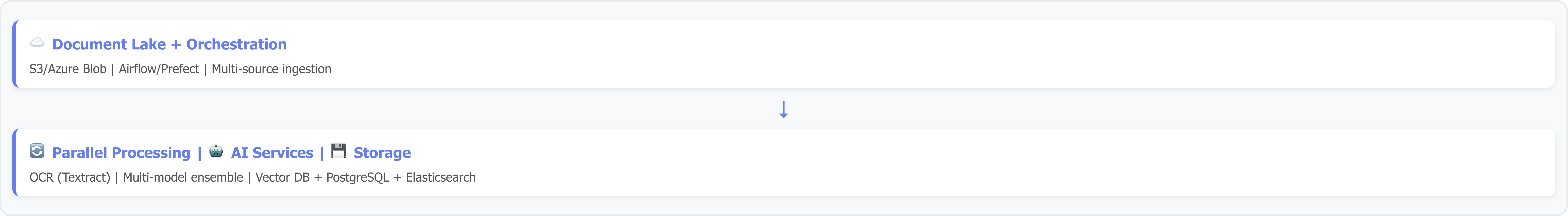
Modularity & Scalability Features

- ✓ **Pluggable Components:** Swap LLM, storage, or OCR engines
- ✓ **Parallel Processing:** Handle multiple documents simultaneously
- ✓ **Horizontal Scaling:** Ready for worker pools and distributed processing
- ✓ **Extensible Schema:** Add new document types by defining schemas

🚀 Beyond POC: Production Architecture

Enterprise-Grade Document Intelligence Platform

🏗️ Production Architecture



☁️ **Document Lake + Orchestration**

S3/Azure Blob | Airflow/Prefect | Multi-source ingestion

↓

🔄 **Parallel Processing** | 🤖 **AI Services** | 💾 **Storage**

OCR (Textract) | Multi-model ensemble | Vector DB + PostgreSQL + Elasticsearch

🎯 Downstream Applications (Real Use Cases)

📊 Financial Reporting

$200.21 CAD total processed. Export to Excel/PowerBI for vendor analysis

🔍 Smart Search

"Find all WeWork invoices" → Instant semantic search results

📈 Aggregation

Uber ($64.46) + WeWork ($36.75) + Cargo ($99) → Trends & forecasting

⚖️ Compliance

Track contract expiry dates, payment deadlines, automated alerts

📚 Legal Research

"Pentcho Tchomakov" → Find all related documents, cross-reference

✨ Summarization

"3 invoices totaling $200.21 from Uber, WeWork, Cargo in June-July"

✅ POC Achievements & Extracted Fields

- ✓ **All 4 Required Fields:** Client names, amounts, dates, involved parties
- ✓ **95% accuracy** across all 3 documents
- ✓ **Multiple formats:** JSON + CSV + Pandas
- ✓ **Modular design** ready for production scaling
- ✓ **Zero API costs** - 100% local processing
- ✓ **Fast:** ~21 seconds (3 documents)

💾 **Output Files Generated**