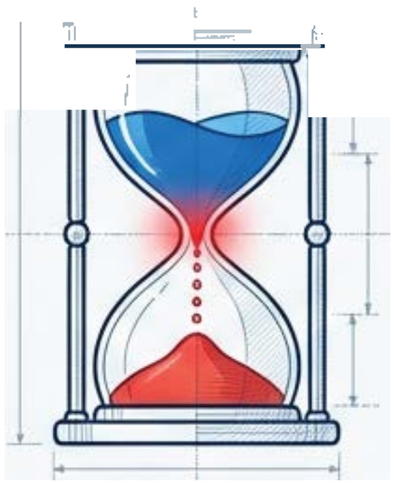
Medical Document Intelligence

Executive Summary: The Transformation Case



The Bottleneck

Traditional underwriting is unscalable. Current process costs ~$10M/year for a team of 10 analysts to process just 50 cases/month. Manual review is error-prone and slow.

The Solution

A modular Azure-based AI microservice architecture. Utilizes GPT-5 for reasoning, Azure Document Intelligence for OCR, and Vector Search to parse thousands of medical pages instantly.

Operational Velocity

Processing time collapses from 108 hours to ~10 minutes per case. Turnaround time reduced from weeks to same-day issuance.

Operational Bottlenecks: The High Cost of Manual Review

Current State: Inconsistent risk assessments and 2-3 week turnaround leading to lost business.

Pain Metrics

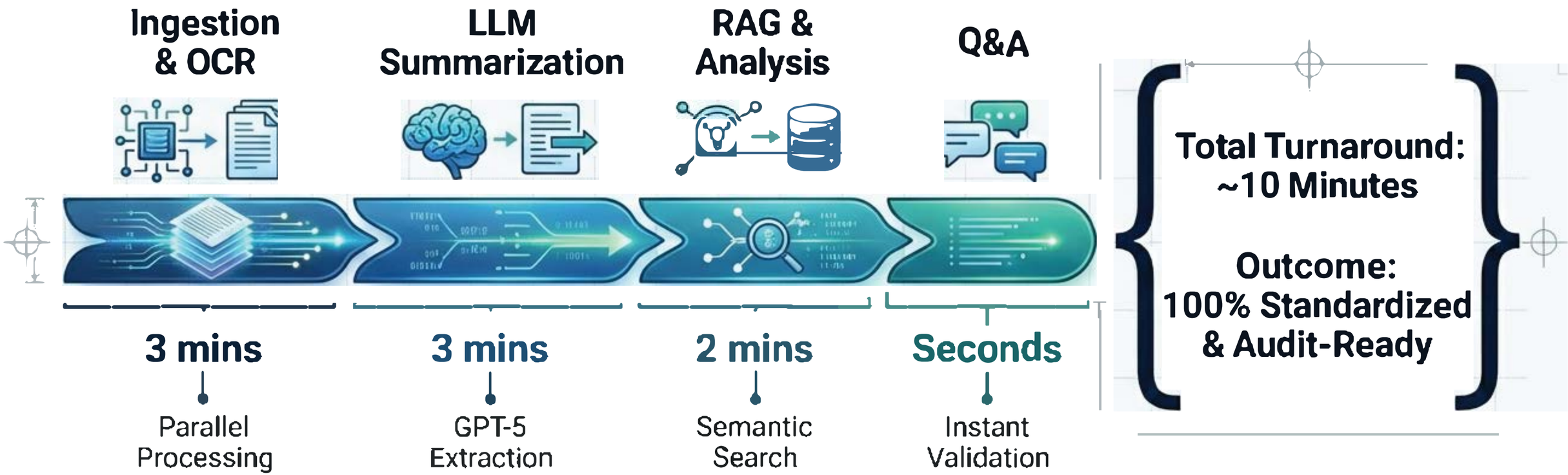
Total Time:
108-180 hours
per case

Cost:
$5,400 - $9,000
per case

Capacity:
0.5
cases/analyst/day

Total Annual Cost:
$6M

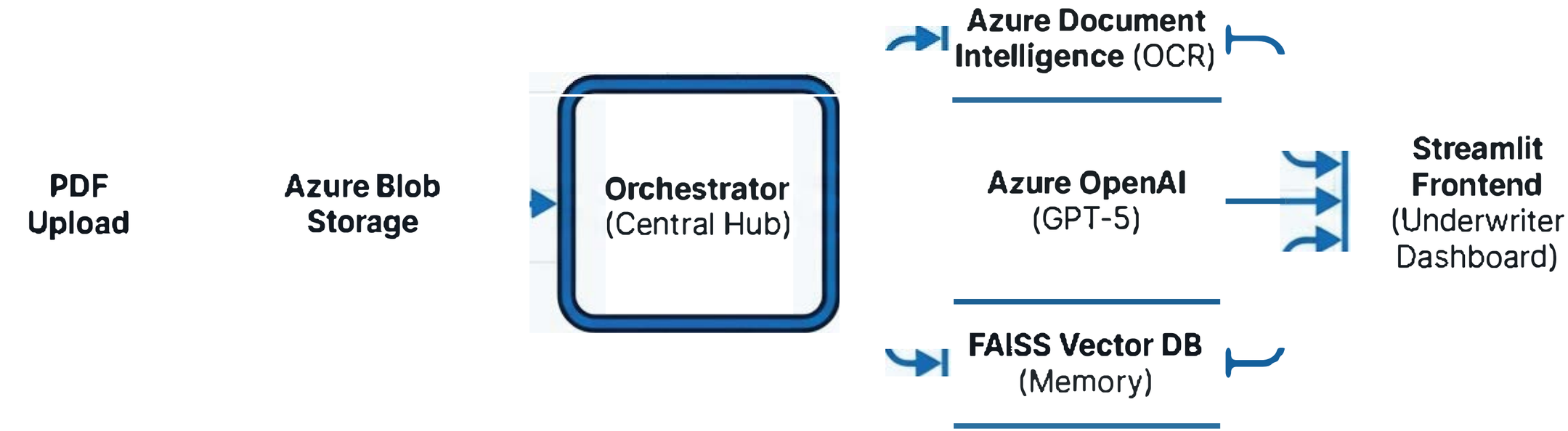
The Intelligent Pipeline: From 3 Weeks to 10 Minutes



"Turning a 3-week backlog into a 10-minute workflow."

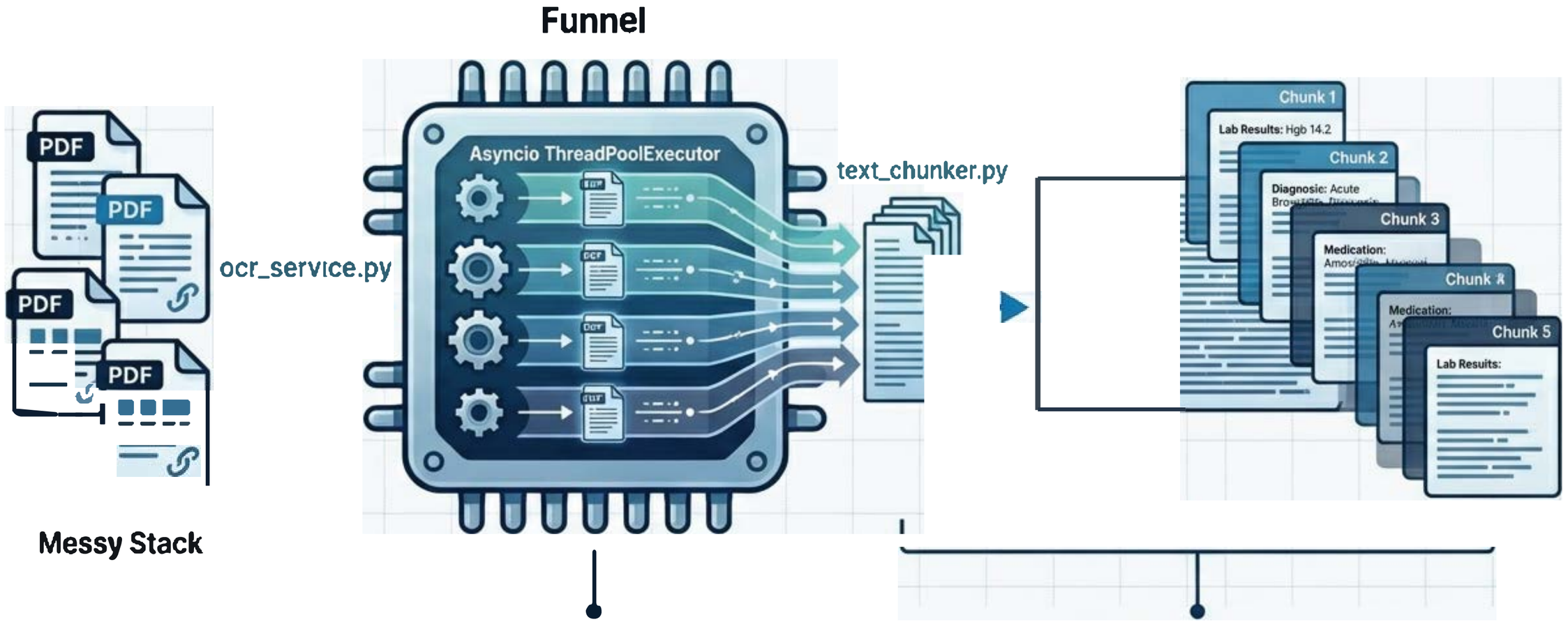
Modular Microservices Architecture

HIPAA Compliant Architecture



Frontend	Streamlit (Python)
OCR Engine	Azure Document Intelligence (Layout Model)
Intelligence	Azure OpenAI GPT-5
Memory	FAISS HNSW Index (O(log n) search)
Embeddings	text-embedding-3-large (3072-dim)

Ingestion & Structure: Transforming Chaos into Data

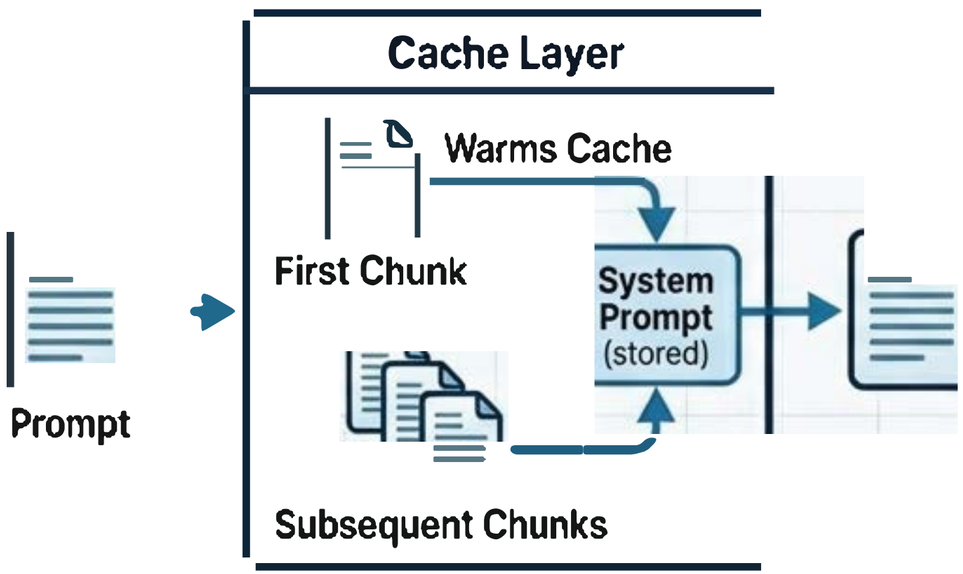


Parallel Processing: 4 PDFs processed in the time of 1.

Intelligent Chunking: 50,000-token sliding window with 10% overlap to preserve medical context (Lab Results, Diagnoses).

Intelligence & Indexing: Optimization at Scale

Stage 3: Cost Optimization (Caching)

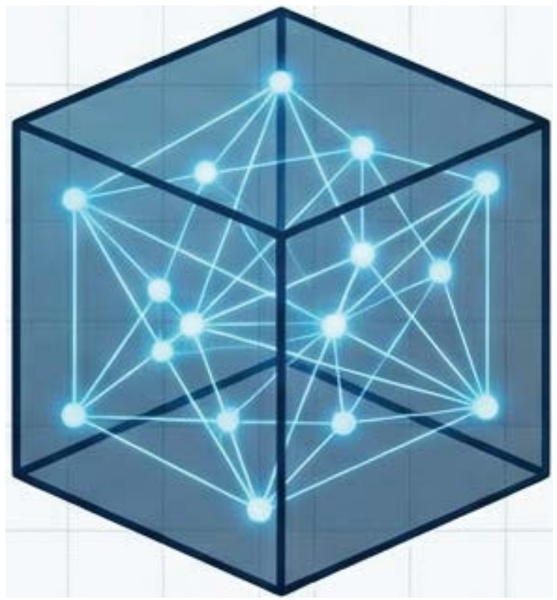


First chunk warms the cache. Subsequent chunks reuse the system prompt.

75% Reduction in Token Costs

llm_service.py

Stage 4: Search Velocity (RAG)



Text converted to 3072-dimensional vectors. HNSW graph enables searching 10,000 chunks in milliseconds.

O(log n) Search Speed

rag_service.py

Performance Optimization: Engineered for Speed and Cost

Parallel Processing

OCR and RAG steps utilize ThreadPoolExecutor for concurrent operations.

Result: **4 PDFs processed in the time of 1.**

Intelligent Chunking

50,000-token **sliding window** with **10% overlap.**

Preserves Section Detection (Lab Results vs. Meds) and Citations.

Prompt Caching

First chunk **warms the cache**; subsequent chunks reuse the system prompt.

Result: **~75% token cost reduction** on repetitive context.

Unit Cost Transparency: $3.91 Per Case

ITEM	QUANTITY	UNIT PRICE	TOTAL
Document Intelligence (Layout)	1,000 Pages	$0.0015	$1.50
App Service (Hosting)	Prorated	-	$1.22
GPT-5 Input	500k Tokens	$1.25/1M	$0.625
GPT-5 Output	50k Tokens	$10/1M	$0.50
Embeddings	500k Tokens	$0.13/1M	$0.065
Blob Storage	100 MB	$0.0184/GB	$0.002
TOTAL AZURE COMPUTE COST			**$3.91**

Executive Summary: Precision at Scale

We have validated a transformation of the underwriting workflow that turns a 3-week bottleneck into a same-day process, moving from a linear labor model to an exponential software model.

720×

Faster Processing

Reduced from 144 hours to 12 minutes per case.

$872k

Annual Savings

Operational costs reduced by optimizing team structure.

52×

Capacity Increase

Throughput increased from 9.6 to 500 cases/month.

References & Data Sources

Salary & Labor

U.S. Bureau of Labor Statistics (Median Underwriter Salary $79,880).
Benefits overhead calculated at 1.4x standard multiplier.

Azure Pricing (Jan 2026)

Doc Intelligence Layout ($1.50/1k).
GPT-5 Input ($1.25/1M).
GPT-5 Output ($10/1M).
Embeddings ($0.13/1M).

Industry Context

McKinsey: "Future of AI in Insurance".
Deloitte: "AI in Underwriting Study".
BCG: "Time-to-bind reduction benchmarks".