Report: How the Extraction Layer Works in the Side-by-Side PDF Viewer

1. Introduction

The extraction layer is the foundational engine behind the interactive functionality of the side-by-side PDF viewer. It transforms raw PDFs into structured metadata, text chunks, normalized bounding boxes, and embeddings. These outputs allow the viewer to highlight exact locations, synchronize scrolling, ground form fields, and enable semantic search through RAG (Retrieval-Augmented Generation).

2. Purpose of the Extraction Layer

The extraction layer prepares PDF documents for intelligent consumption by: - Extracting text and layout information from PDF pages. - Producing normalized bounding boxes for visual highlighting. - Segmenting text into semantic chunks for precision retrieval. - Generating embeddings for semantic search and AI reasoning. - Storing metadata in the data lake, SQL database, and vector database.

3. Extraction Workflow Overview

The extraction workflow progresses through several phases: 1. PDF Upload: The raw PDF is stored in the data lake. 2. Parsing & Layout Extraction: Text blocks and coordinates are gathered. 3. Normalization: Coordinates are normalized for browser rendering. 4. Chunking: Text is segmented into meaningful units. 5. Embedding Generation: Chunks are converted into semantic embeddings. 6. Metadata Persistence: All structured data is written to storage backends.

4. How the Viewer Uses Extraction Outputs

The side-by-side PDF viewer relies heavily on extraction data: - Highlight Rendering: Normalized bounding boxes allow precise overlays. - Scroll-to-Location: Page numbers and bounding boxes guide navigation. - Semantic Search: Embeddings enable concept-level search. - RAG Integration: The assistant can quote and highlight source locations. - Form Field Grounding: Each field can link to exact PDF text.

5. Efficiency Benefits

The extraction layer improves performance and system scalability by: - Eliminating expensive real-time parsing. - Allowing metadata caching for instant viewer load. - Supporting batch embedding and parallel processing. - Storing results in optimized formats (SQL + Vector DB + Parquet). - Enabling incremental updates when documents change.

6. End-to-End Lifecycle in the Viewer

Once extraction is completed, the viewer workflow looks like this: 1. Viewer loads metadata from SQL (fast, cached). 2. User interacts with form, chat, or search. 3. Viewer retrieves the matching chunk. 4. Chunk metadata provides page and bounding box. 5. Viewer scrolls to the exact PDF location and highlights it. This creates a smooth, intelligent, fully synchronized document experience.

7. Conclusion

The extraction layer is essential for converting static PDFs into interactive, AI-powered documents. It provides the structural and semantic data that makes the side-by-side viewer precise, efficient, and capable of advanced retrieval. Without the extraction layer, the system would lack grounding, highlighting, and semantic understanding, reducing it to a simple PDF reader rather than an intelligent workspace.