Complete Workflow — What happens when a PDF is uploaded

1. File Upload & Pre-Processing

Step	Action
1.1	User uploads PDF
1.2	System checks if PDF is digital text or scanned
1.3	If scanned → OCR engine (Tesseract / Azure OCR / AWS Textract) converts to machine-readable text
1.4	Extract metadata (author, pages count, headings if present)

---

2. Document Structuring & Layout Understanding

Step	Action
2.1	Identify Table of Contents (ToC) automatically
2.2	Detect headings, sub-headings, paragraphs using layout parsers (PyMuPDF, PDFPlumber, LayoutLMv3)
2.3	Separate text logically → **Chapter** → **Section** → **Page** → **Paragraph**
2.4	Summarize each section so search result previews are meaningful

✓ No text is cut across chapters
✓ Hierarchy is preserved

---

3. Data Extraction

Extracted Component	What Happens
🔷 Text	Cleaned, chunked & embedded into vectors
🔷 Tables	Extract using Camelot/Tabula → Convert to JSON/CSV → Store in NoSQL

Extracted Component	What Happens
🔷 Images/Charts	Captured + AI generates captions → Embedded for search

---

4. Storage & Indexing

After extraction, 3 different storage systems are used:

Type	Storage	Purpose
Text	Vector Database (Pinecone, Qdrant, Weaviate, Milvus)	Semantic search, Q&A retrieval
Tables	NoSQL DB (MongoDB, DynamoDB, Elasticsearch)	Exact data querying
Images + Charts	Object storage + vector embeddings	Visual search & reference

---

5. Search Pipeline

User enters a query → Behind the scenes:

1. Query is converted into embedding (vector)
2. Vector DB finds most similar document chunks
3. Table DB is queried if question relates to numeric or structured data
4. Relevant images/charts retrieved using visual embeddings
5. Response is ranked + returned with source link + page number

You can ask:

> "Show policy for leave encashment in HR handbook."

The system returns:
• Page/section text
• Related tables
• Related image/chart summary

---

6. Optional Stretch Add-Ons

Feature	How It Works
Multi-language	OCR + embeddings in multilingual models (LaBSE, MUSE, mBERT)
Handwritten text	Handwriting OCR (Vision API/Azure Read)
Chart data extraction	Computer vision + graph digitization (PlotDigitizer, ChartOCR)

Final Output

Output	Description
Structured JSON (Full document mapping)	
Vectorized text for semantic Q&A	
Parsed tables for precise query search	
Image/chart metadata for visual lookup	
Search interface UI	
