RAG Document Ingestion & Retrieval — Tools & Component Mapping Note

1. Purpose

Map **system components to tool categories** in a way that is **defensible, replaceable, and industry-aligned**.

This document answers: - *What kind of tools are used where? - Why those tools fit that component? - What is critical vs replaceable?*

No tool is assumed permanent.

---

2. Guiding Principles for Tool Selection

1. **Components choose tools, not the other way around**
2. **Authoritative data must live in stable systems**
3. **ML tools are optional and replaceable**
4. **Retrieval tools optimize speed, not truth**
5. **Every external system must reference SQL IDs**

---

3. Component → Tool Mapping

3.1 Document Registry

Responsibility - Register documents - Track lifecycle status

Tool Category - Relational Database

Typical Choices - PostgreSQL / MySQL

Why - Strong consistency - Schema enforcement - Easy auditing

Criticality: 🔴Critical **Replaceable**: 🟢Yes (any RDBMS)

3.2 Object Storage (Supporting Component)

Responsibility - Store original PDFs - Store extracted images & OCR artifacts

Tool Category - Object Storage

Typical Choices - S3 / GCS / Azure Blob

Why - Cheap - Immutable - Scales independently

Criticality: 🔴Critical **Replaceable**: 🟢Yes

3.3 Probe Engine

Responsibility - Structural inspection - Sampling pages

Tool Category - PDF parsing utilities

Typical Choices - PDFMiner - PyMuPDF - Apache Tika

Why - Fast - Deterministic - No hallucination risk

Criticality: 🟡Medium **Replaceable**: 🟢Yes

3.4 Logical Decision Engine (Control Plane)

Responsibility - Rule evaluation - Strategy selection

Tool Category - Rule engine / application logic

Typical Choices - Plain application code - Config-driven rule evaluation (YAML/JSON)

Why - Transparency - Easy tuning - No ML dependency

Criticality: 🔴Critical **Replaceable**: 🟡Partially (rules must exist)

3.5 Parsing Engine (Data Plane)

Responsibility - Deterministic extraction - ML/OCR fallback

Tool Categories - Deterministic parsers - OCR engines

Typical Choices - Tabula / Camelot (tables) - PDF text extractors - Tesseract / Textract (OCR)

Why - Pipeline-first minimizes risk - ML used only as repair mechanism

Criticality: 🔴Critical **Replaceable**: 🟢Yes (tool-by-tool)

3.6 Quality & Trust Evaluator

Responsibility - Confidence scoring - Risk classification

Tool Category - Heuristic scoring + lightweight ML (optional)

Typical Choices - Rule-based scoring - Simple classifiers

Why - Predictability - Auditable thresholds

Criticality: 🔴Critical **Replaceable**: 🟡Logic must remain

3.7 Promotion Engine

Responsibility - Decide data destinations - Enforce promotion rules

Tool Category - Application logic

Typical Choices - Workflow step - Message-driven service

Why - Centralized enforcement - Prevents leakage into retrieval

Criticality: 🔴Critical **Replaceable**: 🔴No (conceptually)

3.8 Vector Database (Primary Retrieval)

Responsibility - Semantic similarity search

Tool Category - Vector index

Typical Choices - FAISS - Pinecone - Weaviate

Why - Fast retrieval - Scales independently

Criticality: 🟡Medium **Replaceable**: 🟢Yes

3.9 Knowledge Graph (Selective Reasoning)

Responsibility - Relationship-based reasoning

Tool Category - Graph database

Typical Choices - Neo4j - RDF triple stores

Why - Explicit relationships - Explainable reasoning

Criticality: 🟡Conditional **Replaceable**: 🟢Yes

3.10 Retrieval & Query Layer

Responsibility - Answer user queries - Combine vector + graph results

Tool Category - Application service - LLM (optional, bounded)

Why - Orchestration, not truth creation

Criticality: 🟡Medium **Replaceable**: 🟢Yes

---

4. Tool Philosophy Summary

- SQL and object storage are **foundational**
- Vector DB is **primary consumption**, not truth
- Knowledge Graph is **situational**, not universal
- ML tools are **assistive**, never authoritative

---

5. Readiness Statement

With components and tools mapped: - File structure can now be derived cleanly - Services can be split logically - Tool choices can change without redesign

---

End of tools & component mapping note.