**High-Level Design: Document Indexing & RAG System**

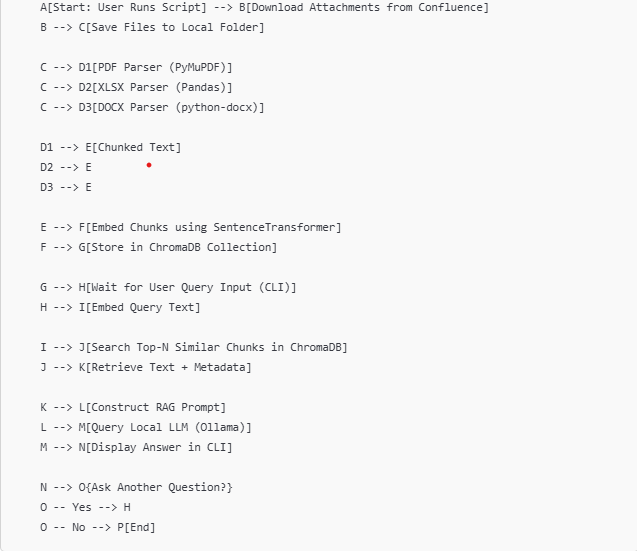
**Objective**

To build a system that:

* Downloads documents from Confluence
* Extracts and chunks data from PDF, DOCX, and XLSX files
* Embeds data using SentenceTransformers
* Stores and queries embeddings in ChromaDB
* Uses a local LLM (via Ollama) for contextual question-answering (RAG)

**Data Flow Diagram**

****

****

**Tech Stach Summary:**

| **Component** | **Tool/Library** |
| --- | --- |
| Vector DB | ChromaDB |
| Embedding Model | SentenceTransformers |
| LLM | Ollama (phi:latest) |
| File Parsing | PyMuPDF, pandas, python-docx |
| HTTP API | requests + HTTPBasicAuth |
| Local Storage | OS filesystem |
| Interface | Command Line (CLI) |