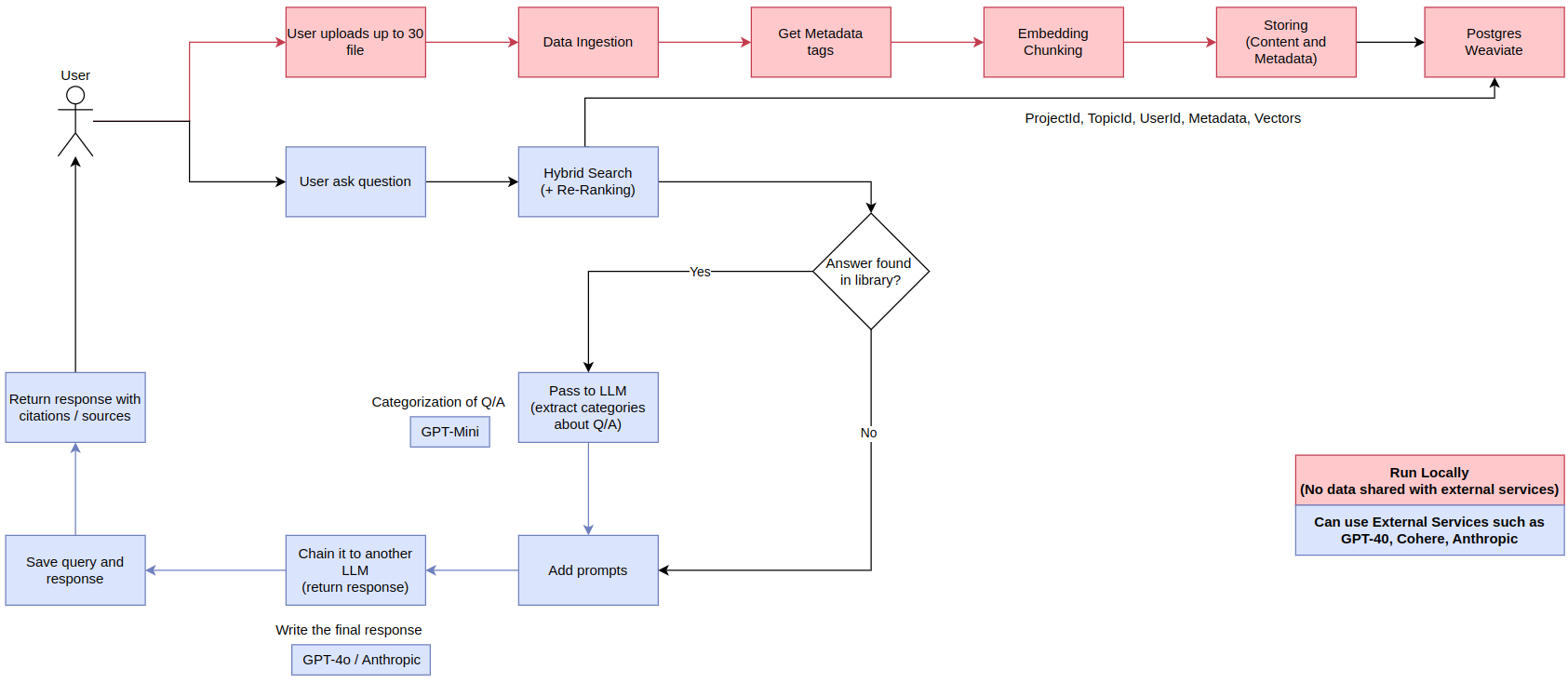
The RAG application needs to work with French and English for now. The application needs to be running on Docker. Data ingestion pipeline needs to be fully running locally so no data sharing with external models. Our server is 32GB/8vCPUs. High Level overview:



Details about the file types (<https://docs.unstructured.io/platform/supported-file-types>). We want to have all the file types supported by the Unstructured library (in addition to images and videos) and below is how the expected behavior for each of these files. We are usually focused on PDFs, Epub, Microsoft and Open Office files, videos, Images and HTML files.

**Main Process**

We start with the process that is needed for all documents. So for all file types we need to do the below

- Content extraction

- Named Entity Recognition

- Categorization (5 top categories that explain the document),

- Chunk the extracted content

- Embed chunks

- Store and Retrieve using hybrid search

The below files require **additional processing**

For PDF Files (I’ll send them separately)

Extract images, tables and text content and if the pdf is:

Fully Structured PDF (like a book): Extract author information, chapters, sections, and page numbers

Semi-structured PDF (like a scientific article): Extract author information and chapter when available

Unstructured PDF (Text/Images/Tables): No metadata extraction (only NER, categorization as mentioned in the main process)

\* Some PDFs might need to be OCR

Document Files (Microsoft/Open Office - Word, PDF, EPUB)

Extract images, tables, and text content and page numbers (if possible sections as well)

Spreadsheets (Microsoft/Open Office - Excel)

Extract content organized by tab names

Presentations (Microsoft/Open Office - PowerPoint)

Extract data and their slide numbers

HTML Files or fetch data from websites

Extract all content including images

Videos

Usually the videos tend to be technical. If the video contains presentations, diagrams, text or tables we would like to have transcription + the content on the screen.

Which frames to OCR/extract the data from?  
Extract a snapshot whenever a frame appears for at least 4 seconds in the video.

Deal with the images/frames as per the description in the images section

Images

Extract content from images specially and add metadata about the content of the image

Have that content embedded by a vision model to understand a bit the context and not only the text

Give you an idea about simple questions that the users could ask

“What does this image talk about”

"What does this book discuss?"

"What are the references used by the author?"

"What does Chapter 1 cover?"

"In which chapter does the main character first appear?"

"What does the conclusion discuss?"

"What does each part of the book focus on?"

"How does Chapter 3 relate to Chapter 7?"

"Which chapters cover technology topics?"

“Which slide discusses the technology?”

“What is the content of the first tab in Excel?”

“what's the lecture in the video about?”

Some additional prompts which require searching across different files or all of them. Assuming we have the below file types in a single topic/project:

* meeting\_notes.txt
* financial\_report.docx
* research\_paper.pdf
* product\_image.jpg
* sales\_data\_chart.png
* customer\_data.xlsx
* marketing\_presentation.pptx
* webpage\_content.html
* data\_export.json
* meeting\_recording.mp4 / mp3

The user can ask to:

* Summarize the key findings from these documents
* What does chapter 2 of the research paper talk about?
* Combine the data from this spreadsheet with the visual information from the images and create a report
* Find all mentions of 'X project' across these files and return the results
* Extract all the text from these PDFs and Word documents
* Identify the main topics discussed in meeting notes?
* Extract all the images from the pdf and word documents
* Based on these documents and images, what are the potential risks and opportunities
* Create a table summarizing the information from these documents
* What does the financial report talk about?
* What happened in the presentation in the meeting recording?