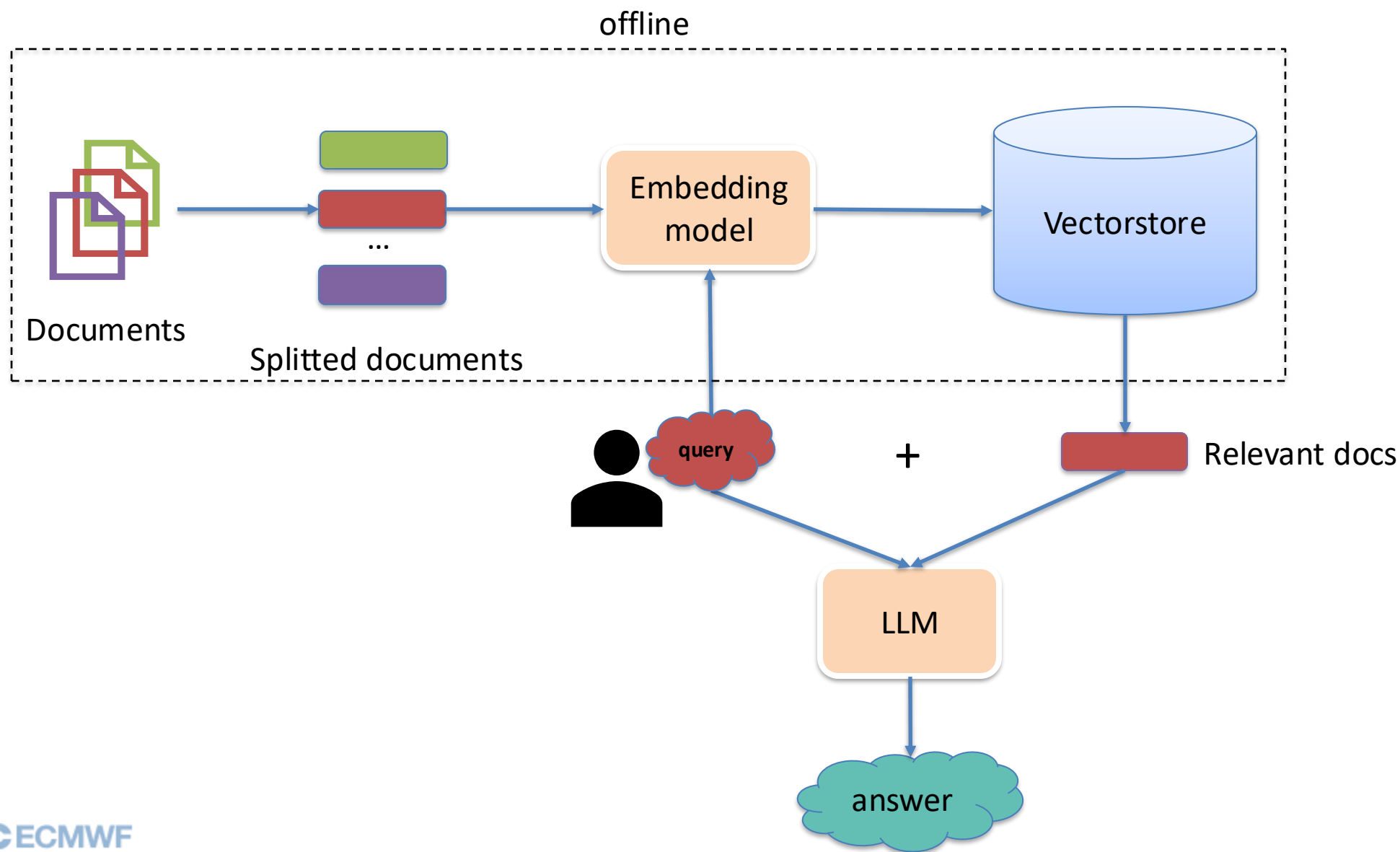
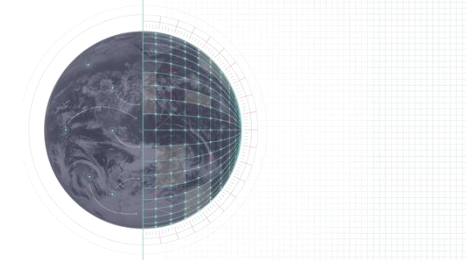
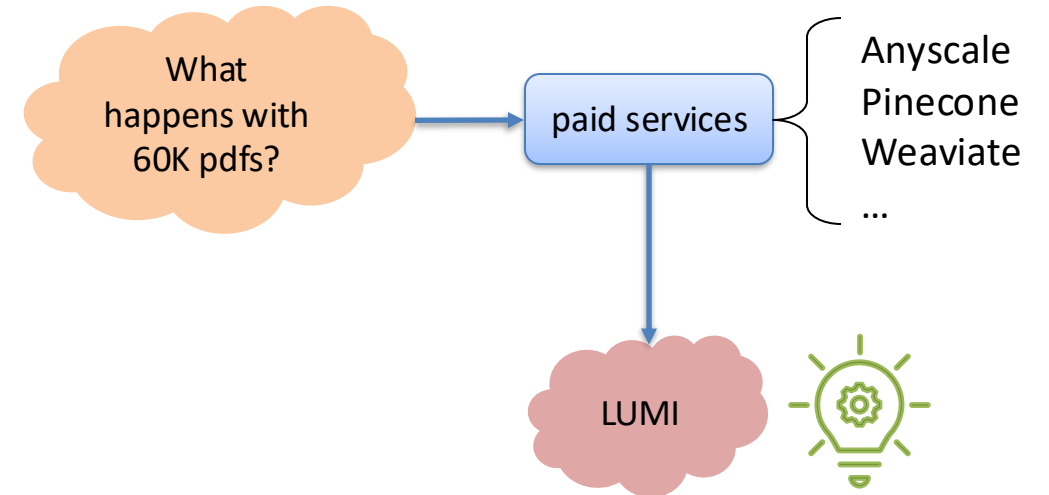
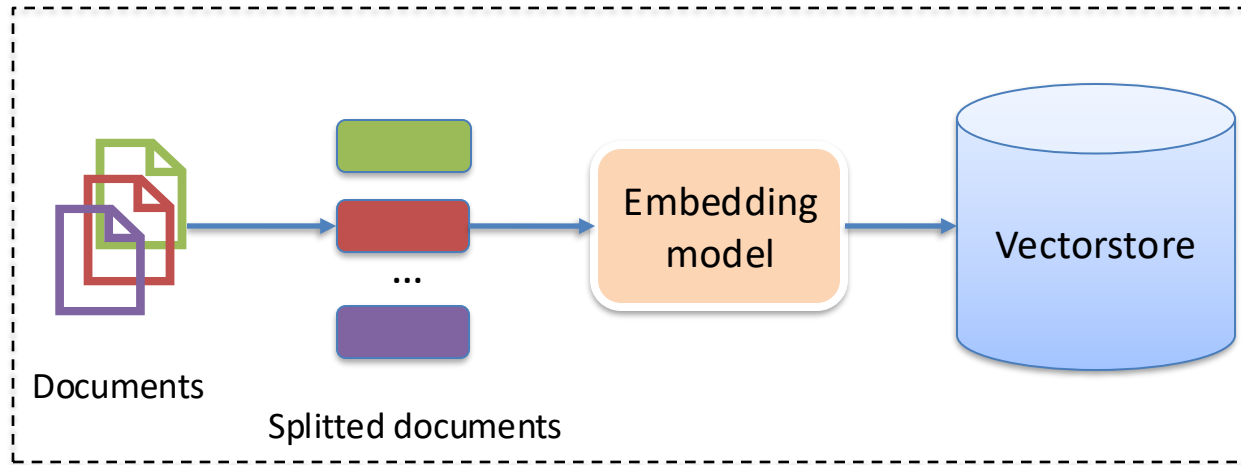
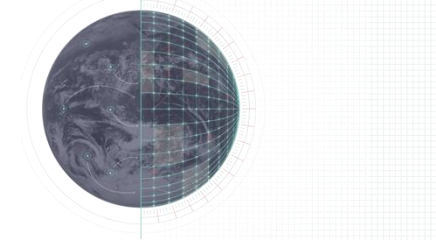


RAG PIPELINE



BUILD VECTOR STORE

- Langchain



```
documents = PyMuPDFLoader(pdf_path).load()
```

```
text_splitter = RecursiveCharacterTextSplitter(chunk_size = 1000, chunk_overlap = 200)
```

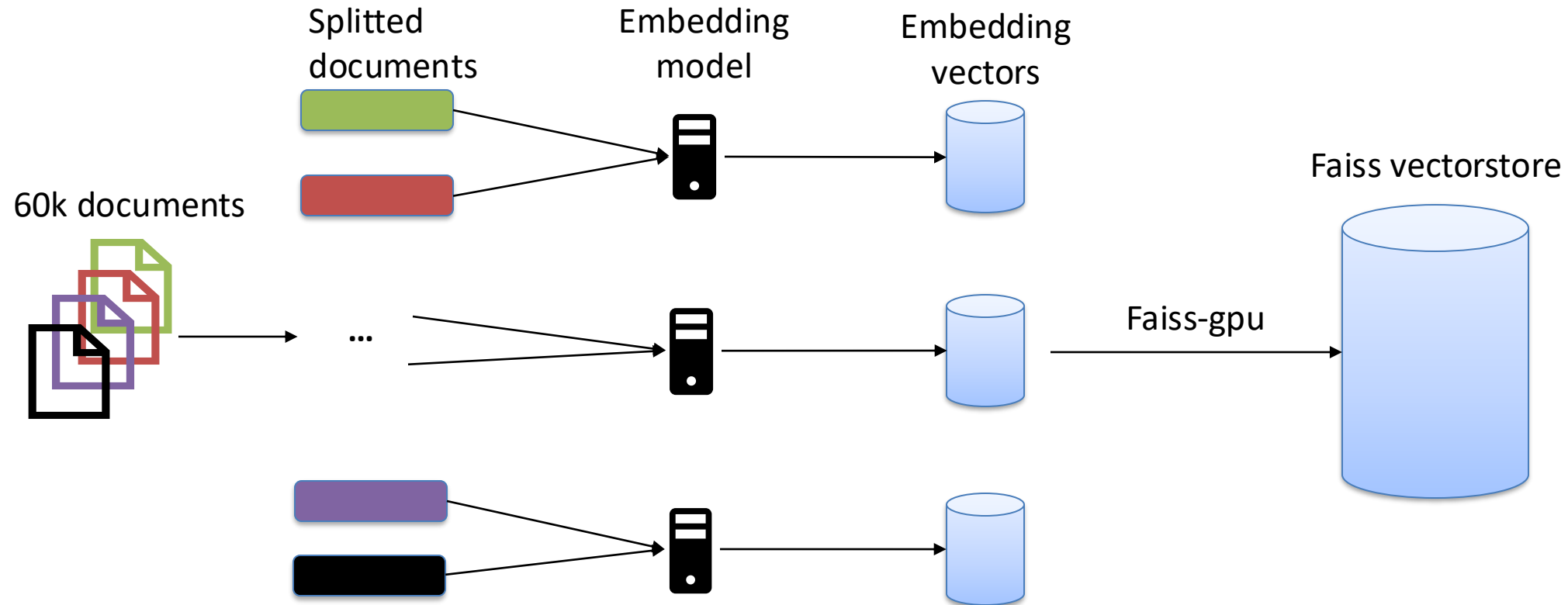
```
split_documents = text_splitter.split_documents(documents)
```

```
vectordb = Chroma.from_documents(split_documents, embedding = OpenAIEmbeddings(), persist_directory = "./db")
```

BUILD VECTOR STORE FROM SCRATCH



- PyTorch Distributed Data Parallel (DDP)

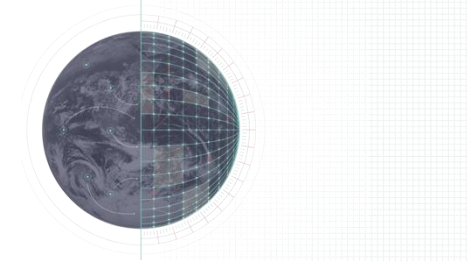


```
sampler = DistributedSampler(dataset)
model = DDP(model, device_ids = [local_rank])
```

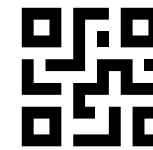
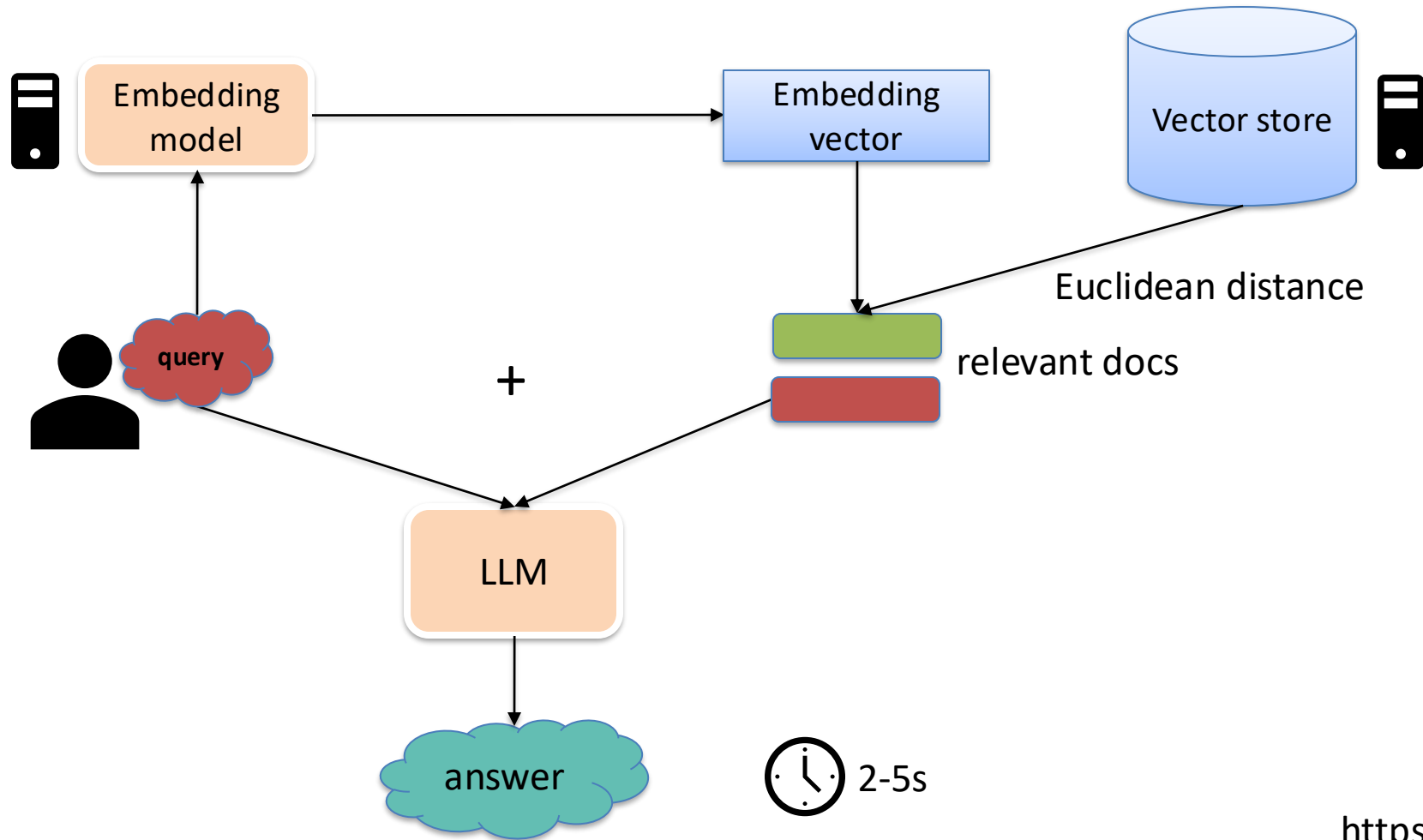
⌚ ~1B tokens
15-45 minutes

```
gpu_index = faiss.index_cpu_to_all_gpus(cpu_index)
```

⌚ 3.3 million vectors
0.5s



RETRIEVAL & GENERATION



<https://github.com/CSCfi/RAG-60K>