CIKLUM

RAG Architecture & Design Patterns

October 2025

Speaker

Lucian Gruia

Data Science Principal Lead

- Head of Engineering, Romania
- 13+ years of experience in Software Engineering
- Led cross-functional teams across 8 time zones in Telecom, Fintech, Aerospace, and MedTech
- Mentor @Google Developer Groups Romania
- Research Assistant, AI Multimedia Lab (National University of Science & Technology, Bucharest)
- Supervised 12 BSc and MSc diploma projects

CIKLUM

Agenda

01 Data Chunking

02 Advanced Retrieval

03 History Management

04 Reasoning & Context Awareness

05 Coding RAG

CIKLUM

Data Chunking, Advanced Retrieval

Data Chunking and LLMs

CIKLUM

LLMs also have a limited capacity for context.

Just as humans **cannot digest unlimited context**, these models have a specific size limit for the content they can process.

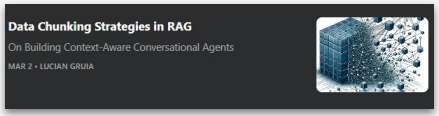
So, what about situations involving very large amounts of data?

Consider a specific use case, such as a book. It's too large to pass the entire book as **the context** for the current prompt, so it **needs to be divided** before being stored in the database.

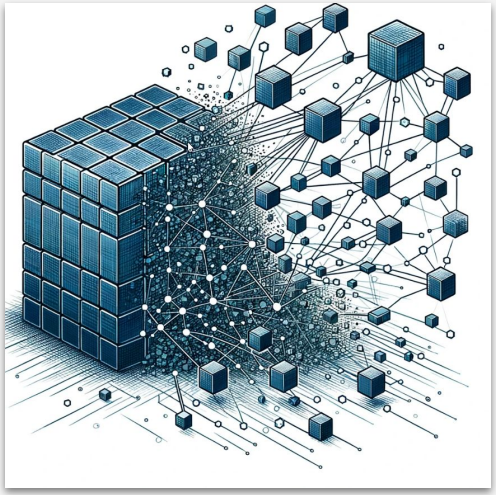
This process is known as **data chunking**.

Types of Data chunking (by size):

- Fixed-size
- Variable Chunking
- Semantic Chunking



Read more: Data Chunking Strategies in RAG

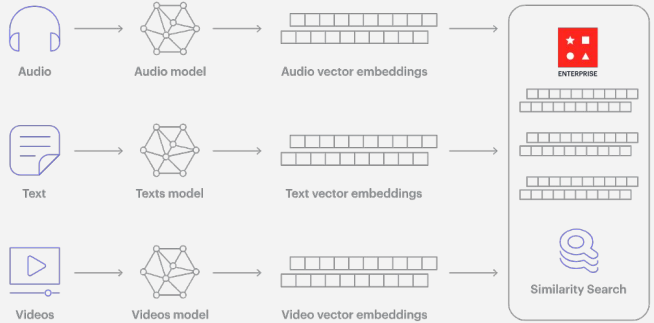


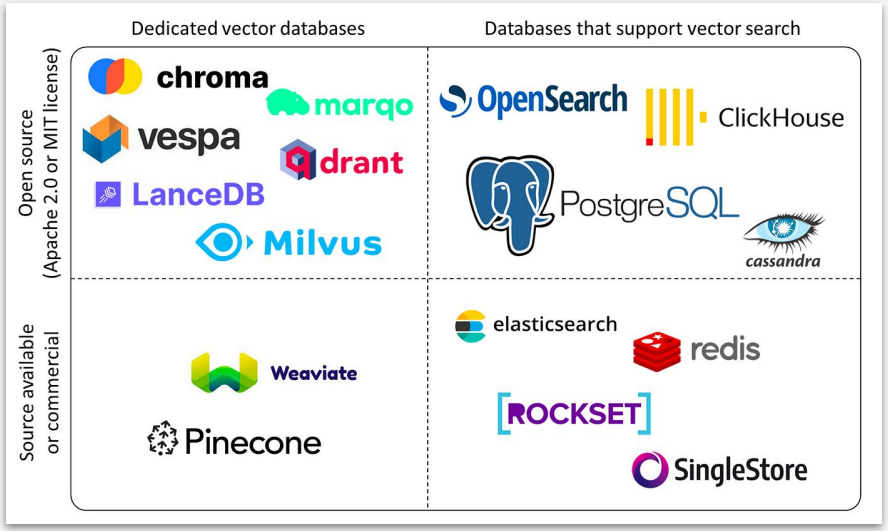
Generated with DALL·E 3

Vector Databases

CIKLUM

- **Store, index, and manage high-dimensional vector data (aka Embeddings)**
- **Perform similarity search**
  - Cosine similarity
  - Pearson Correlation
  - Levenshtein Distance
  - Jaccard Similarity
  - Euclidean distance
  - Dot Product
  - Minkowski Distance





Images source: Datacamp - The Top 5 Vector Databases

CIKLUM

Advanced Retrieval

Retrieval

The retrieval component is responsible for sourcing relevant information from a database or document collection based on the user's query or input. This information is attached to prompt and passed to the LLM as an **enriched context**.

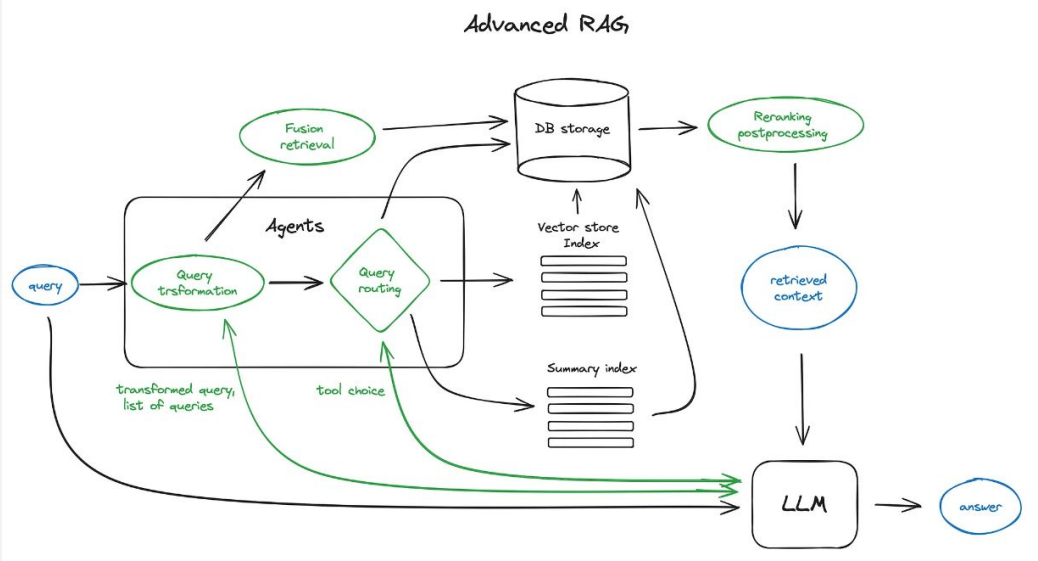
Techniques

PERFORM RETRIEVAL

- Vector Store Index
- Hierarchical Indices (summaries associated)
- Hypothetical Questions and HyDE
- **Context** enrichment
  - Sentence window Retrieval
  - Auto-merging retrieval (aka Parent-Document)
  - Fusion retrieval/hybrid search

POST RETRIEVAL

- Reranking & Filtering
- Query Transformations
  - Step-back prompting
  - Prompt rewriting/Reformulate
- Evaluation



Images sources: TowardsAI - Advanced RAG Techniques: an Illustrated Overview

CIKLUM

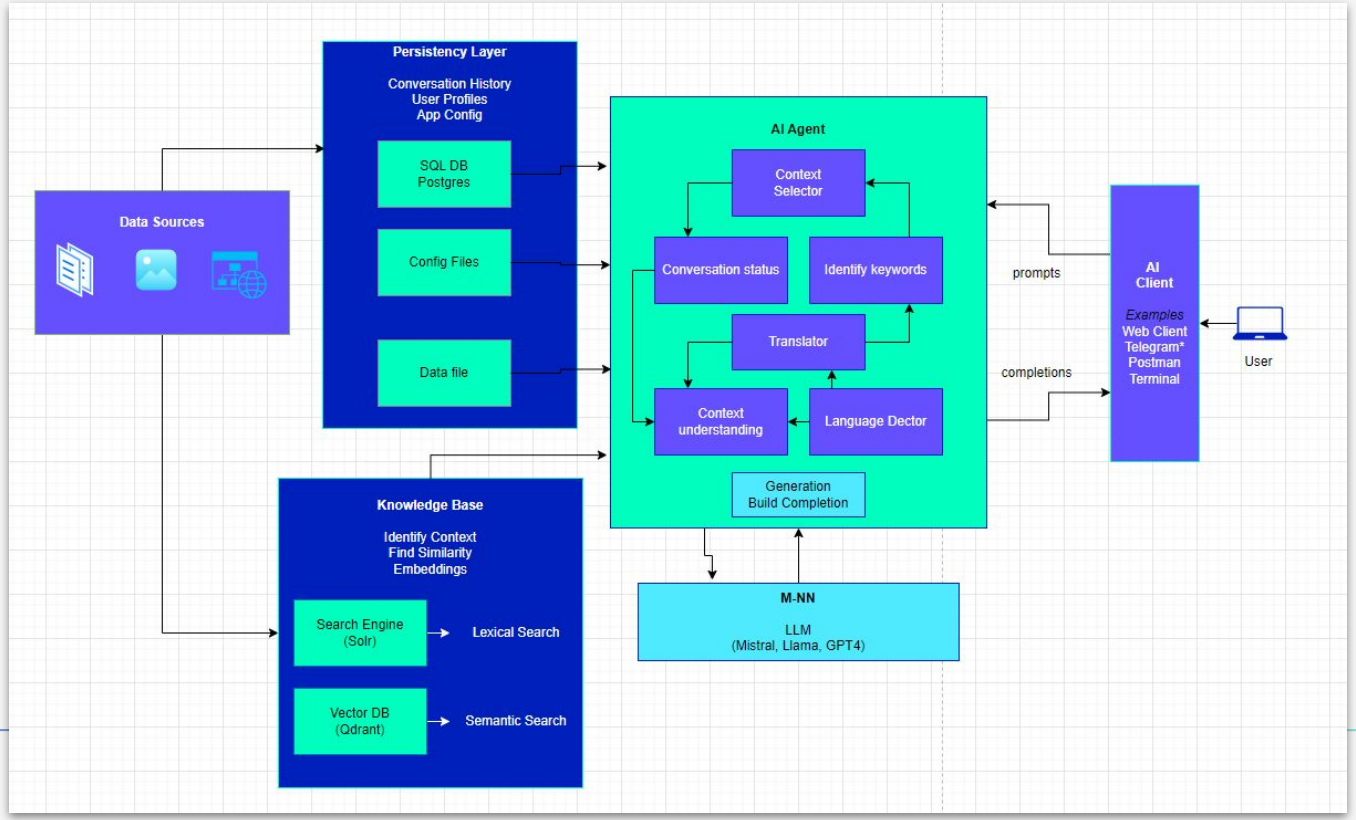
History Management

CIKLUM

Reasoning, Context Awareness

RAG Architecture

CIKLUM



CIKLUM

Coding RAG apps

Thank you!