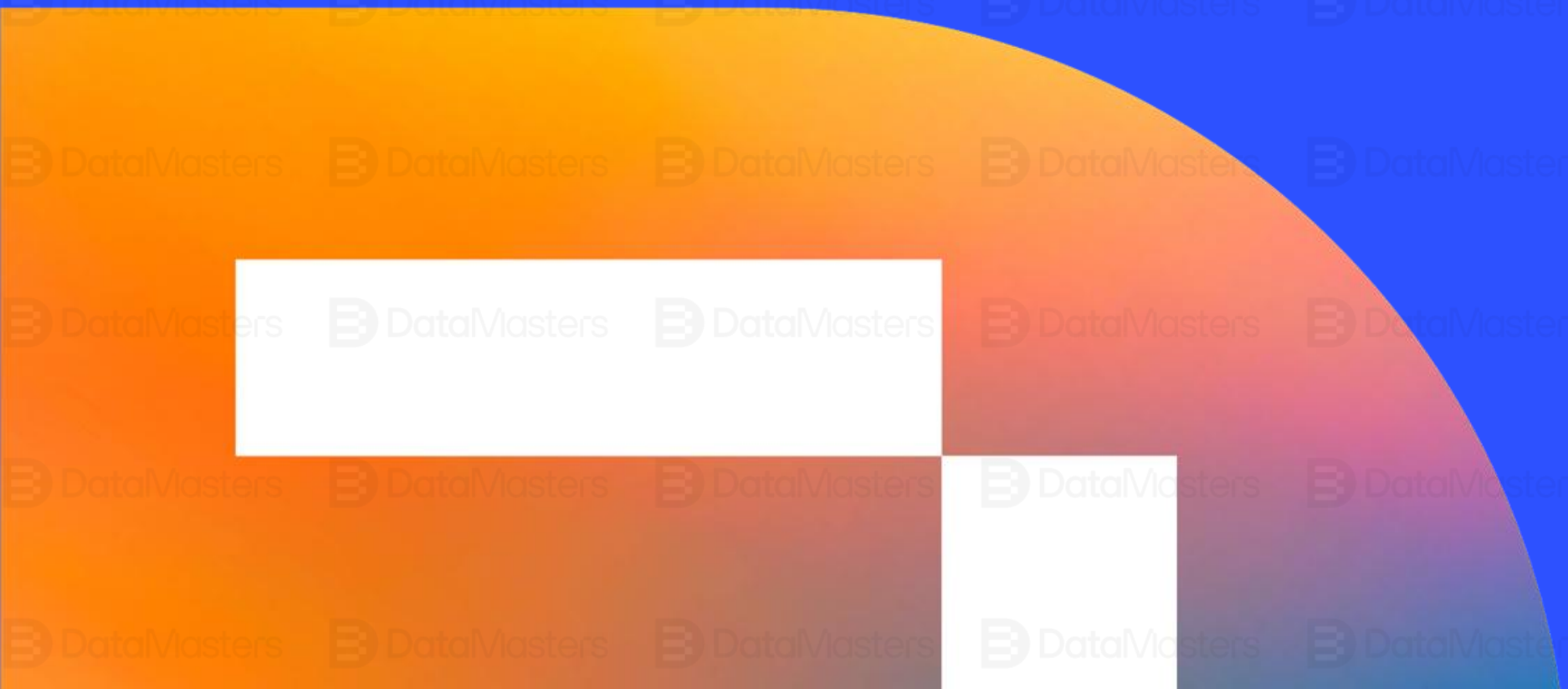
RAG – Retrieval Augmented Generation

DAY 4/7



DataMasters

RAG – Retrieval Augmented Generation

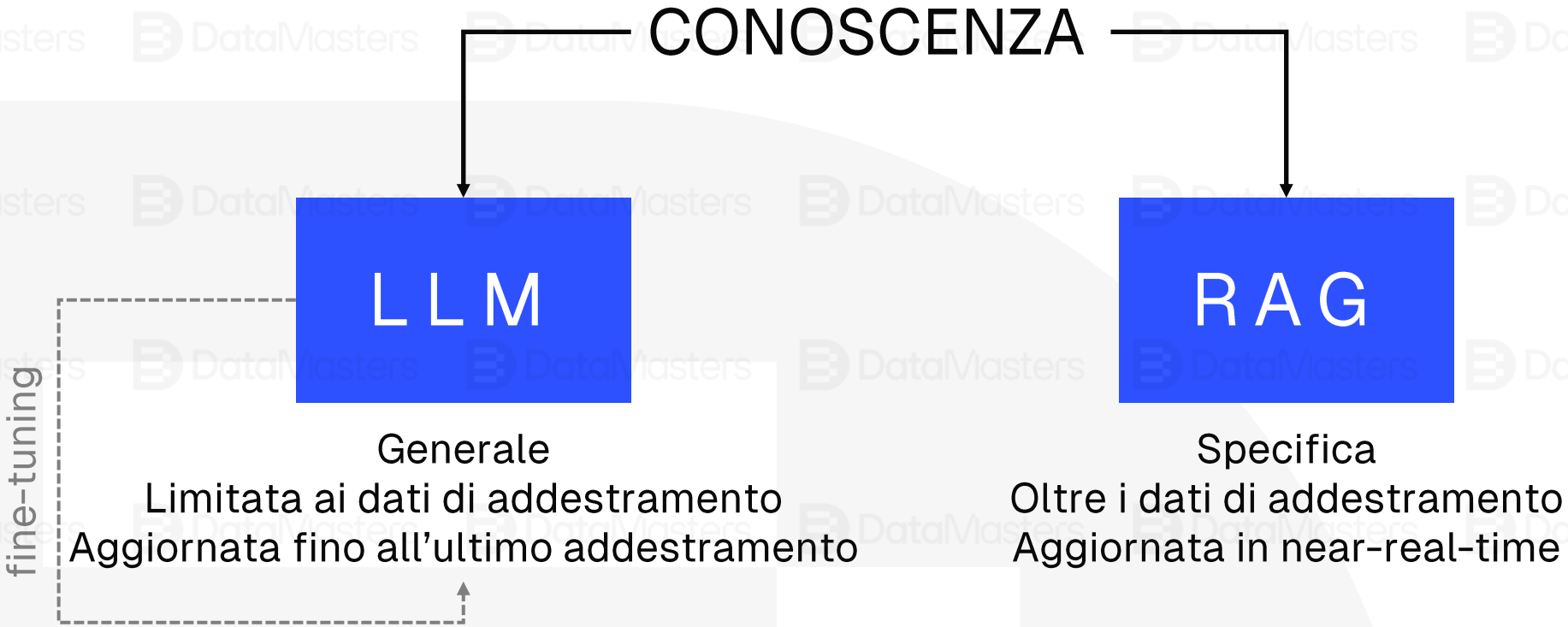
→ Introduzione alle tecniche di RAG

→ Vector-DB

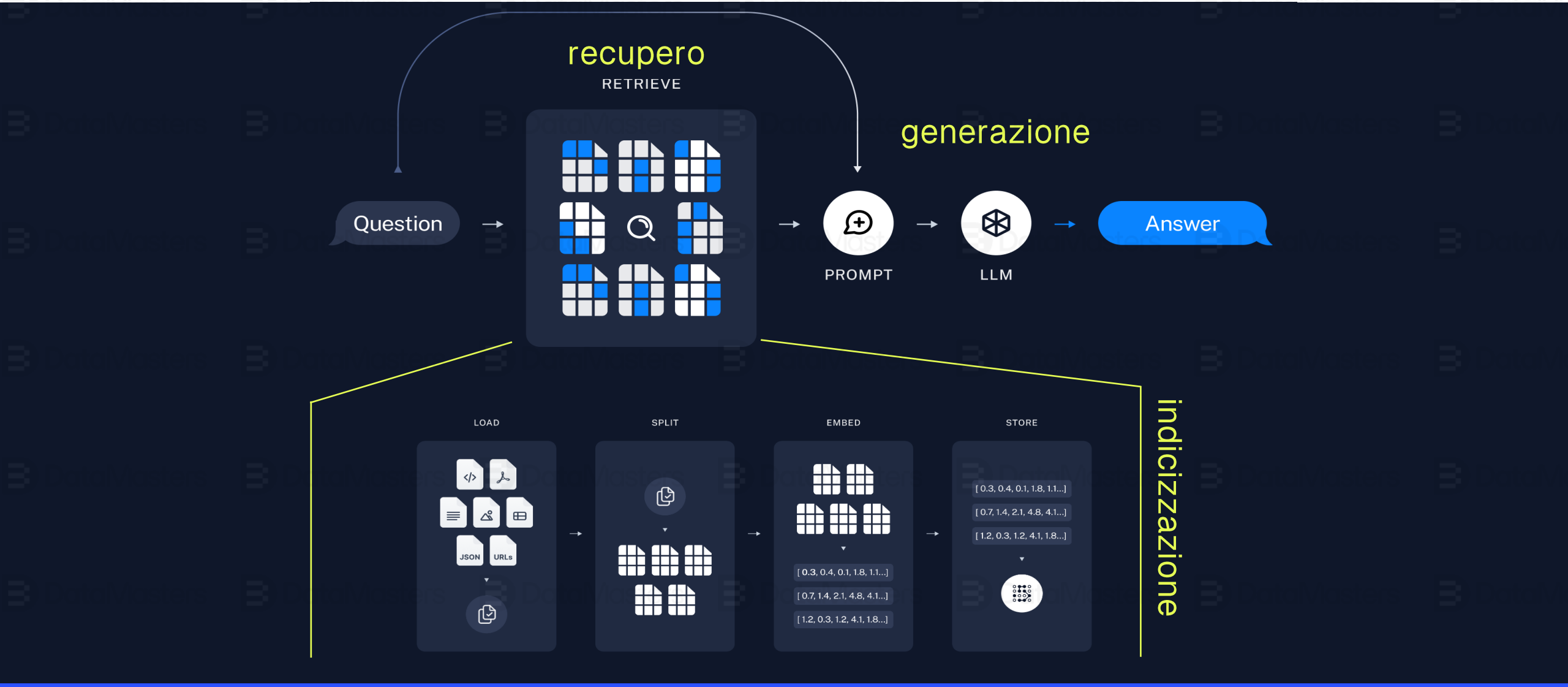
→ Esempi di RAG su documenti e pagine web con:

- OpenAI Embeddings + ChromaDB + OpenAI GPT3.5
- Ollama + Nomic + Facebook FAISS + Google Gemma
- Ollama + GPT4All + ChromaDB + Llama
- RAG strutturato su PDF
- ParentDocumentRetriever

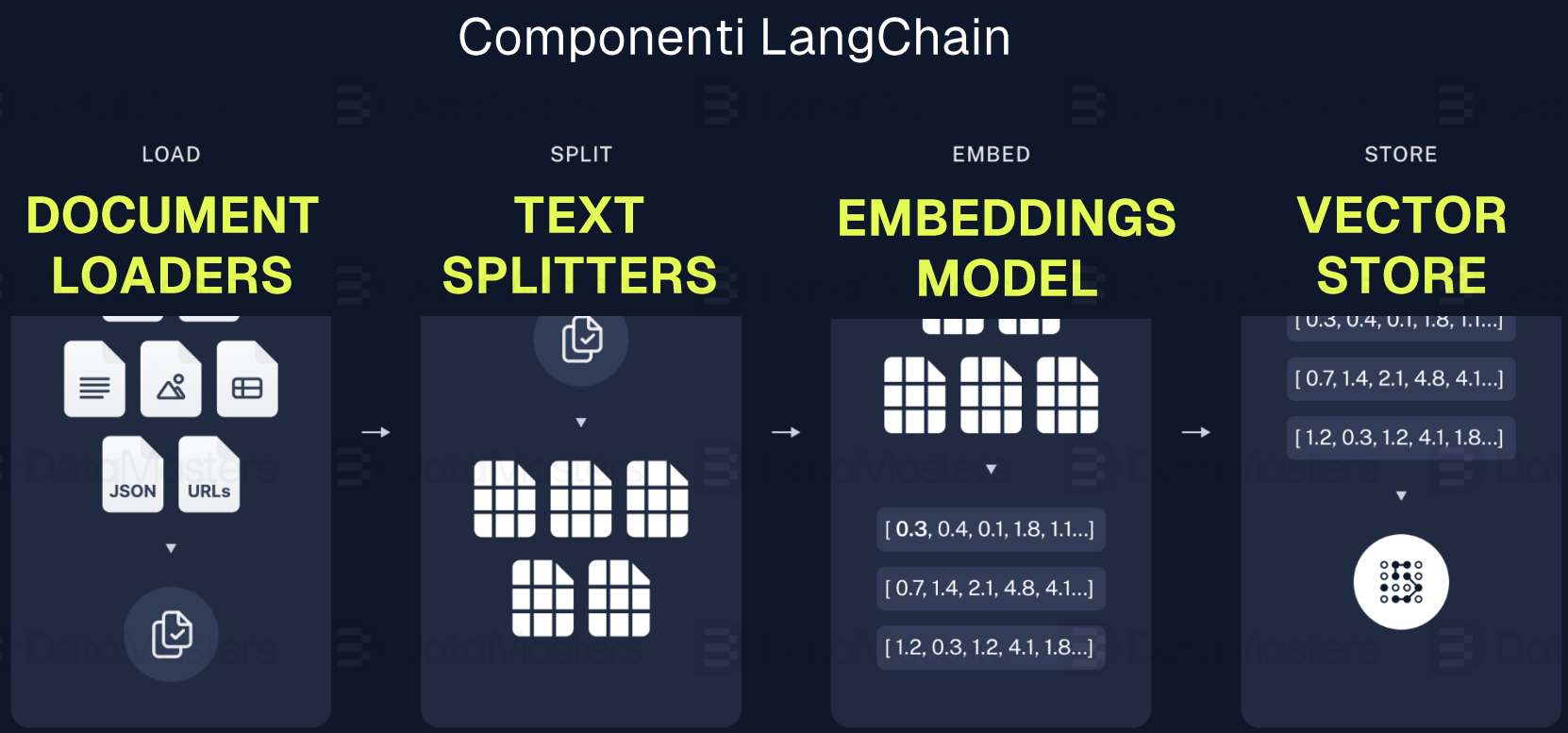
Perché usare RAG



Architettura RAG



Indicizzazione



Vector Database Management System (VDBMS)

→ Collezione di dati memorizzati in forma di vettori

- Permette ai SW di elaborare efficacemente testi

→ Velocità nel trovare elementi simili

- Ricerche
- Raccomandazioni
- Generazione

- Comparazioni
- Identificare Relazioni
- Sfruttare il Contesto

Principali VDBMS

- → Apache Cassandra
- → Chroma
- → Azure Cosmos DB
- → Elasticsearch
- → LlamaIndex
- → ...
- → MongoDB Atlas
- → OpenSearch
- → Pinecone
- → PostgreSQL pgVector
- → Redis Stack

Hands On

- Chroma Embeddings Database, FAISS & VectorDB
- Question Answering
  - da testo
  - dal web
  - da PDF
  - su modelli locali
  - con output strutturato
- ParentDocumentRetriever