INTRODUCTION TO RAG

edquest

What is Retrieval-Augmented Generation (RAG)?

- RAG stands for Retrieval-Augmented Generation
- It helps AI find information before answering
- Gives better and more accurate answers

RAG is a smart way to improve AI. Instead of only using its memory, the AI searches through documents or databases before answering. Just like we use Google to check facts, RAG helps AI do the same!

👉 It’s helpful when you want your AI to talk using your own proprietary knowledge base.

info@edquest.co.in

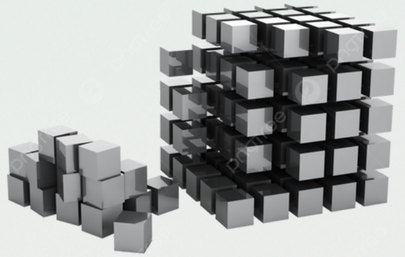
+91-9467752177

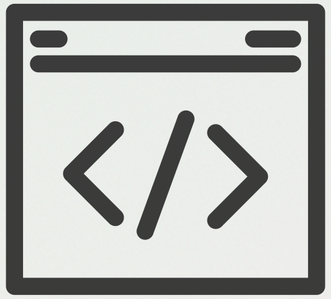
edquest

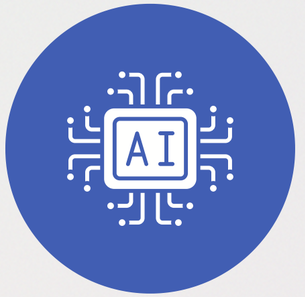
Components of RAG

edquest

- Chunking
- Embeddings
- Vector Databases
- Semantic Search
- Reranking
- LLM
- User Query
- Response Generation







info@edquest.co.in

+91-9467752177

edquest

edquest

WHAT IS CHUNKING?

Chunking: Breaking Big Text into Small Parts

- Break big files into smaller sections (chunks)
- Each chunk should have complete meaning
- Helps AI find the right part of text

Chunking is like cutting a big book into pages. If we break the content properly, AI can search and understand better. If chunks are too big or too small, AI might miss important details.

👉 A good chunk size is usually between 200 to 500 words depending on the use case.

info@edquest.co.in

+91-9467752177

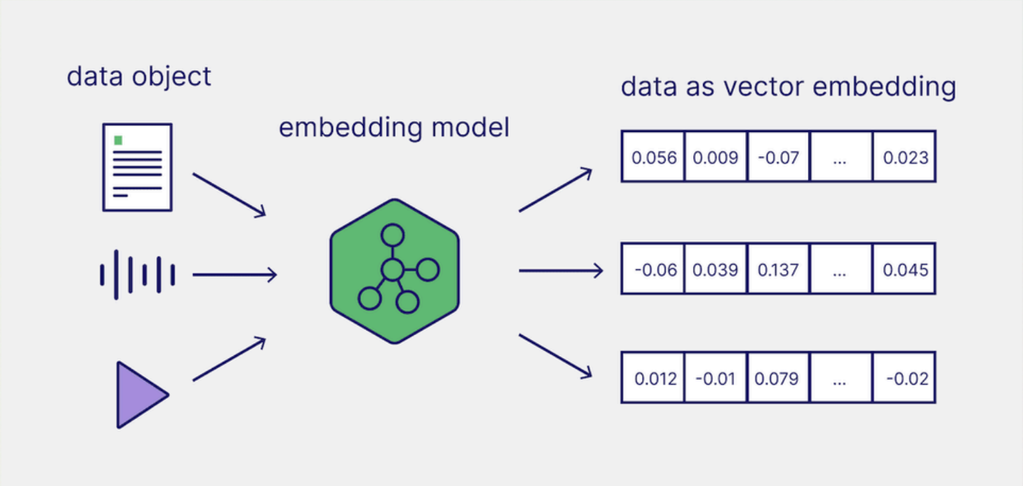
edquest

WHAT IS EMBEDDING?

edquest

AI doesn’t understand plain text the way humans do — it converts text into numerical representations called embeddings.

- Embeddings are numerical representations of text
- Similar meanings → similar vectors
- Enables AI to search by meaning, not just keywords
- Good embeddings capture both context and meaning

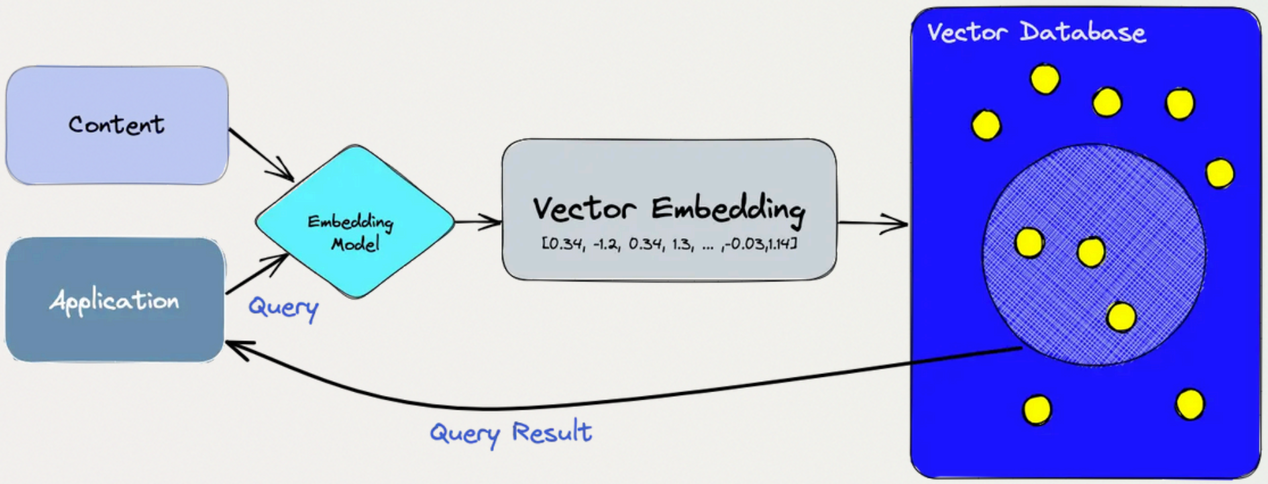


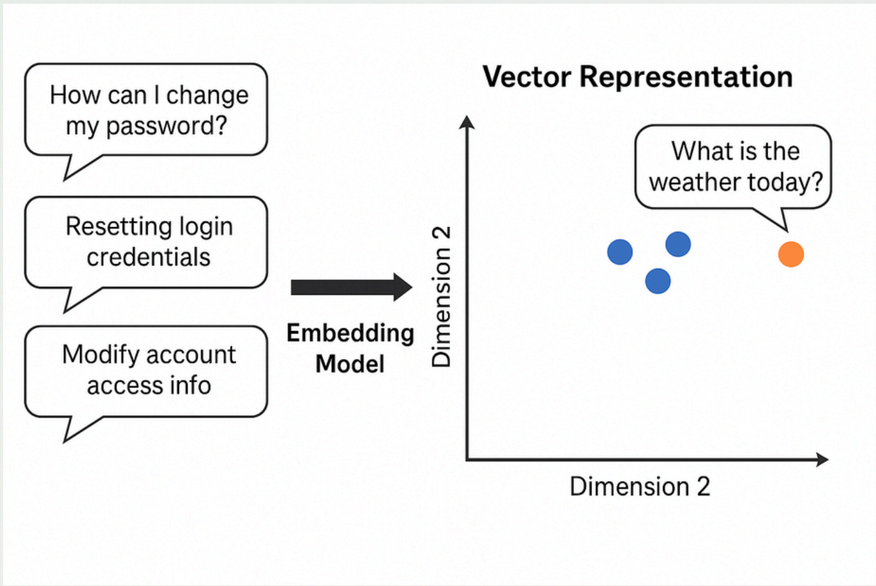
info@edquest.co.in
+91-9467752177
edquest

HOW ARE EMBEDDINGS GENERATED?

edquest

Embeddings are numerical representations of text (or other data) generated by AI models that capture the context and meaning.





NEED FOR A VECTOR DATABASE

edquest

As AI advances, searching by meaning—not just keywords—is essential. That’s where Vector Databases come in.

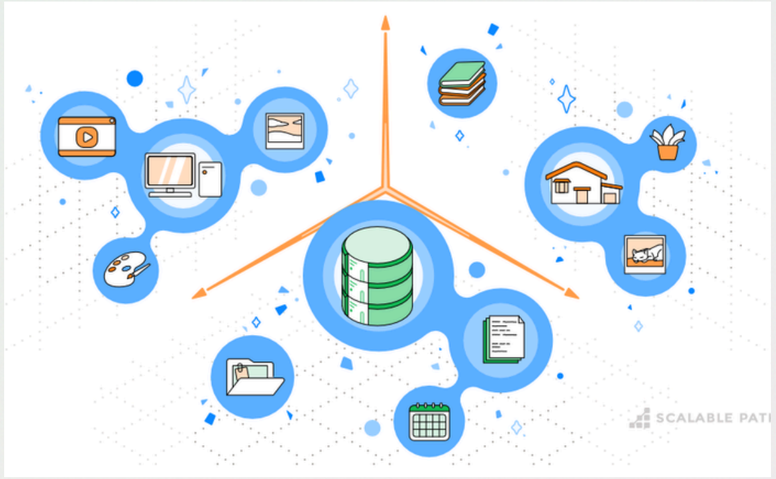
- Traditional DBs can't handle high-dimensional vectors efficiently
- AI models create embeddings that need to be stored and searched
- Essential for recommendations, semantic search, and chatbots
- Vector DBs enable fast, scalable similarity search across millions of embeddings

WHAT IS A VECTOR DATABASE (VECTOR DB)?

edquest

A Vector Database stores and manages embeddings (numerical vectors) to enable search by meaning rather than exact matches.

- Stores vector representations of text, images, or other data
- Enables semantic search — finds results based on similarity of meaning
- Ideal for use with AI models to handle recommendations, Q&A, clustering, and more
- Optimized for high-speed similarity search in large datasets



info@edquest.co.in

+91-9467752177

edquest

KEY CONCEPTS OF VECTOR DATABASES

edquest

Understanding the core ideas behind how Vector DBs work:

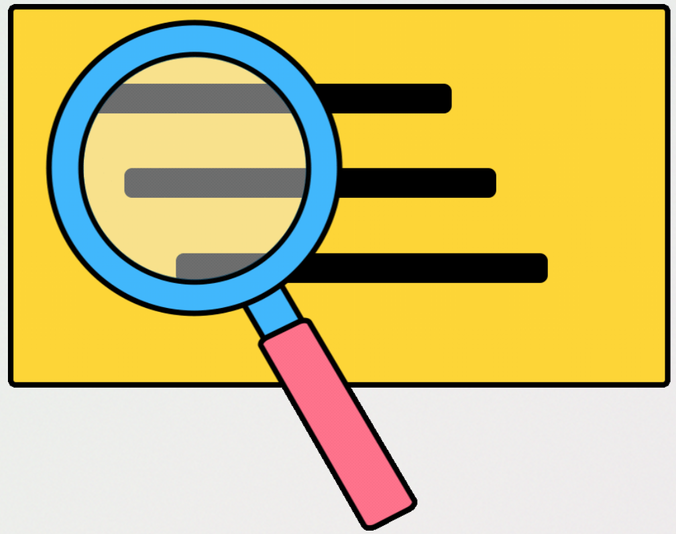
- Embeddings.
- Vector Similarity.
- Indexing.
- Similarity Search (k-NN).
- Scalability.

WHAT IS SEMANTIC SEARCH?

edquest

Semantic Search is a search technique that focuses on the meaning behind the words, not just the exact keywords.

- Understands Meaning, Not Just Words
- Goes Beyond Keyword Matching
- Finds Smarter, More Relevant Results
- Powered by AI and Vector Search
- Enhances Search in Smart Applications



info@edquest.co.in

+91-9467752177

edquest

HOW SEMANTIC SEARCH WORKS

edquest

How It Works:

- Turn Text into Vectors with Embeddings
- Save Those Vectors in a Vector Database
- Convert the User’s Question into a Vector
- Search for Similar Vectors (Like Using Cosine Similarity)
- Get Results That Match Meaning, Not Just Words

info@edquest.co.in

+91-9467752177

edquest

ASKING A QUESTION (USER QUERY)

edquest

- User asks a question
- Question is also converted into a vector
- AI searches for similar chunks in the database

When someone asks a question, the system changes it into a vector (just like chunks). Then it compares this with all saved chunks to find the closest matches.

👉 The quality of the answer depends on how well the question is converted into an embedding.