RAG – RETRIEVAL-AUGMENTED GENERATION

🎯 What this lesson is about:

You'll learn what RAG is, how it makes AI assistants smarter, and how to build a RAG-powered workflow in n8n to generate accurate answers from real-time or internal data.

🔍 What is RAG?

- **RAG = Retrieval + Generation**
- First, it **retrieves real data** (from documents, APIs, databases)
- Then, it **generates a response** based on that data using an LLM
- This prevents hallucinations and gives real, updated answers

📦 Why use RAG?

- ChatGPT and other LLMs have **fixed training knowledge**
- They can't access your internal company data or documents
- RAG enables them to **search external sources** and respond with facts
- Great for dynamic info like refund policies, stock prices, etc.

⚙️ How does RAG work?

1. User asks a question
2. LLM identifies what info is needed
3. RAG **retrieves relevant data** from internal docs or APIs
4. LLM uses that data to generate an answer
5. Result: **Accurate, contextual, up-to-date response**

🔧 RAG Assistant in n8n – Key Components:

- **Tools Agent** – The assistant's brain with system instructions
- **Chat Model** – OpenAI model that generates the final response
- **Postgres Chat Memory** – Stores previous user interactions
- **Supabase Vector Store** – AI-searchable memory from documents
- **Embeddings Node** – Converts text into vectors for semantic search

RAG – RETRIEVAL-AUGMENTED GENERATION

✅ What does RAG enable?

- Real-time answers from internal data
- Smarter, context-aware automation
- Assistants that don’t guess – they fetch the correct answer

🔥 Real-World Use Cases:

- Customer support – Answers based on actual policies, past tickets
- Chatbots – Pull latest updates (news, weather, inventory)
- Education & research – Retrieve the most relevant material
- E-commerce – Stock info, personalized recommendations
- Finance – Up-to-date interest rates, account info
- Healthcare & insurance – Policy details or patient info
- Legal & compliance – Latest laws or contract terms