Video: Video u47GtXwePms Q&A;
URL: https://www.youtube.com/watch?v=u47GtXwePms&t;=126s

Q1: Explain RAG in breif
A1: RAG stands for Retrieval-Augmented Generation. It's a solution pattern that combines a large language model (LLM) with a retrieval system to generate answers to user questions. The retrieval system finds relevant content related to the question, and the LLM uses that content to generate a more accurate and informative response.

Q2: How does RAG works
A2: RAG (Retrieval-Augmented Generation) is a solution pattern used to leverage large language models (LLMs) with your own content. Here's how it works: 1. **Content Chunking**: The content is broken down into smaller chunks, such as paragraphs or pages. 2. **Vectorization**: Each chunk is sent to an LLM, which converts it into a vector (a series of numbers) that represents the essence of the chunk. Similar chunks on similar topics will have similar vectors. 3. **Question Vectorization**: When a user asks a question, it is also sent to the LLM, which converts it into a vector (the question vector). 4. **Comparison**: The question vector is compared mathematically to the vectors of the content chunks to find the top matching chunks (typically the top 5). 5. **Retrieval**: The top matching chunks are retrieved and used to augment the LLM's generation process. 6. **Generation**: The LLM uses the retrieved chunks to generate a response to the user's question. The goal of RAG is to provide a more accurate and relevant response by leveraging the user's own content and using the LLM to generate a response based on the most relevant information. This approach is particularly useful when dealing with large amounts of content and complex topics.

Date Generated: 2025-06-20