Lesson 4: RAG Basics (Retrieval-Augmented Generation)

Objective:

This lesson introduces the concept of Retrieval-Augmented Generation (RAG), a method where external information is retrieved and integrated into the model's responses.

Key Concepts:

- **Retrieval-Augmented Generation (RAG):** A method of combining information retrieval (fetching relevant data from a database or the web) with generative models (like GPT). This allows the model to provide more accurate, context-aware answers by referencing external information.
- **Information Retrieval:** The process of fetching relevant documents or data from a large set.
- **Generative Model:** A machine learning model that generates text, such as GPT-3.

Application:

1. **Building RAG Pipelines:**
    - Combine a retrieval model (e.g., using Elasticsearch or another search engine) with a generative model to improve the responses.
2. **Use Cases:**
    - Improve search results by adding context from generative models.
    - Enhance Q&A systems by pulling in real-time information from the web and integrating it with AI-generated content.

Example:

In a RAG system, a query like “What’s the weather in New York?” could first trigger a search for relevant weather data and then generate a response using this data in combination with the AI’s knowledge