Data Scientist 1 – RAG Challenge

Objective

We want to assess your ability to work with text data, understand the basics of Retrieval-Augmented Generation (RAG), and demonstrate strong problem-solving. This challenge focuses on the fundamentals of building and evaluating a simple RAG pipeline — with cloud deployment as a **bonus**.

The Challenge

You are given two documents (e.g., PDF, docx,). Your task is to:

1. **Data Preparation**
   - Load and preprocess the documents (e.g., cleaning, chunking, text normalisation).
   - Explain your preprocessing decisions.
2. **Retrieval Component**
   - Implement a **basic retrieval step** using either vector search (embeddings) or keyword search
   - Show how a query retrieves the most relevant documents.
3. **Generation Component**
   - Use any LLM interface to combine the retrieved documents with a query.
   - Demonstrate how the LLM uses context from retrieval to answer a question.
4. **Evaluation**
   - Define at least **two evaluation criteria** (e.g., relevance of retrieved docs, accuracy of answers, hallucination rate).
   - Run a few test queries and evaluate performance.
5. **Bonus**
   - Deploy on your preferred cloud platform (e.g. AWS, GCP, Azure)

Deliverables

- GitHub link to Codebase include a ReadMe with your findings & trade-offs for your decisions
- If Bonus work completed, deployed link