**LLM DATA ENGINEER PRE-ASSIGNMENT**

**PLEASE UPLOAD LINK/FILE INTO IQN, IF ABLE TO DO SO – OTHERWISE, PLEASE SEND LINK/FILE DIRECTLY TO MAGNIT**

**ASSIGNMENT IS AS FOLLOWS:**

1. Data Ingestion:
   1. Provide a dataset (e.g., JSON, CSV, or unstructured text files) that includes a mix of structured and unstructured data.
   2. Ask the candidate to create a pipeline to load this data into a database of their choice, ensuring the schema is optimized for querying.
2. Data Preprocessing:
   1. The data may contain noise or require transformation (e.g., text cleaning, parsing nested JSON, handling missing values).
   2. The candidate should demonstrate how they preprocess the data for efficient storage and later retrieval.
3. Vectorization:
   1. Using a pre-trained language model or embeddings model, ask the candidate to convert the unstructured text into embeddings.
   2. Store these embeddings in a vector storage solution of their choice, ensuring the pipeline can handle batch processing for larger datasets.
4. Query and Retrieve:
   1. Create a simple API or script that allows querying based on a given text prompt. The query should retrieve similar embeddings from the vector store and return the corresponding records from the database.
   2. Include a use case for Retriever-Augmented Generation (RAG), where the retrieved data is used to generate a summary or response based on the query.
5. Documentation:
   1. The candidate should document their code, the thought process behind their design choices, and any trade-offs they considered (e.g., schema design, vector storage approach, etc.).
6. Bonus:
   1. Implement monitoring or logging for the data pipeline to track the data flow and identify potential bottlenecks.
   2. Optimize the pipeline for scalability, such as handling larger files or parallel processing.