Assignment 4 - Building a RAG Pipeline with Airflow

Pranjal Mahajan (002375449)

Srushti Patil (002345025)

Ram Srikar Putcha (002304724)

Project Summary:

This project focuses on building an AI-powered information retrieval application to process and query NVIDIA's quarterly reports from the past five years. The system will be designed to efficiently extract and retrieve relevant information from unstructured PDF data using multiple advanced parsing strategies, including custom extraction methods, Docling, and Mistral OCR. The core of the project is a Retrieval-Augmented Generation (RAG) pipeline, which will initially operate with manual embeddings and cosine similarity and later integrate with Pinecone and ChromaDB for advanced retrieval capabilities. The pipeline will support multiple chunking strategies and hybrid search to query specific quarter data. The system will be deployed with Apache Airflow for orchestration, enabling seamless data ingestion, processing, and retrieval workflows. A user-friendly interface will be built using Streamlit, allowing users to upload PDFs, select parsers, choose RAG methods, and interact with the data via FastAPI. Docker containers will be used to deploy both the data pipeline and the user interface, ensuring easy scalability and maintainability.

A screenshot of a computer

AI-generated content may be incorrect.