Project -3: LLMOps
RAG on 10k-reports

This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Problem Statement

Finsights Grey Inc. is an innovative financial technology firm that specializes in providing advanced analytics and insights for investment management and financial planning. The company handles an extensive collection of 10-K reports from various industry players, which contain detailed information about financial performance, risk factors, market trends, and strategic initiatives. Despite the richness of these documents, Finsights Grey's financial analysts struggle with extracting actionable insights efficiently in a short span due to the manual and labor-intensive nature of the analysis. Going through the document to find the exact information needed at the moment takes too long. This bottleneck hampers the company's ability to deliver timely and accurate recommendations to its clients. To overcome these challenges, Finsights Grey Inc. aims to implement a Retrieval-Augmented Generation (RAG) model to automate the extraction, summarization, and analysis of information from the 10-K reports, thereby enhancing the accuracy and speed of their investment insights.

This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Objective

As a Gen AI Data Scientist hired by Finsights Grey Inc., the objective is to develop an advanced RAG-based system to streamline the extraction and analysis of key information from 10-K reports. You are asked to deploy a Gradio app on HuggingFace spaces that can RAG 10-k reports and answer the questions of financial analysts swiftly.

The project will involve testing the RAG system on a current business problem. The Financial analysts are asked to research major cloud and AI platforms such as Amazon AWS, Google Cloud, Microsoft Azure, Meta AI, and IBM Watson to determine the most effective platform for this application. The primary goals include improving the efficiency of data extraction. Once the project is deployed, the system will be tested by a financial analyst with the following questions. Accurate text retrieval for these questions will imply the project's success.

This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Questions to ask the RAG system

We need to ask these questions on the Gradio UI for each company to test the system.

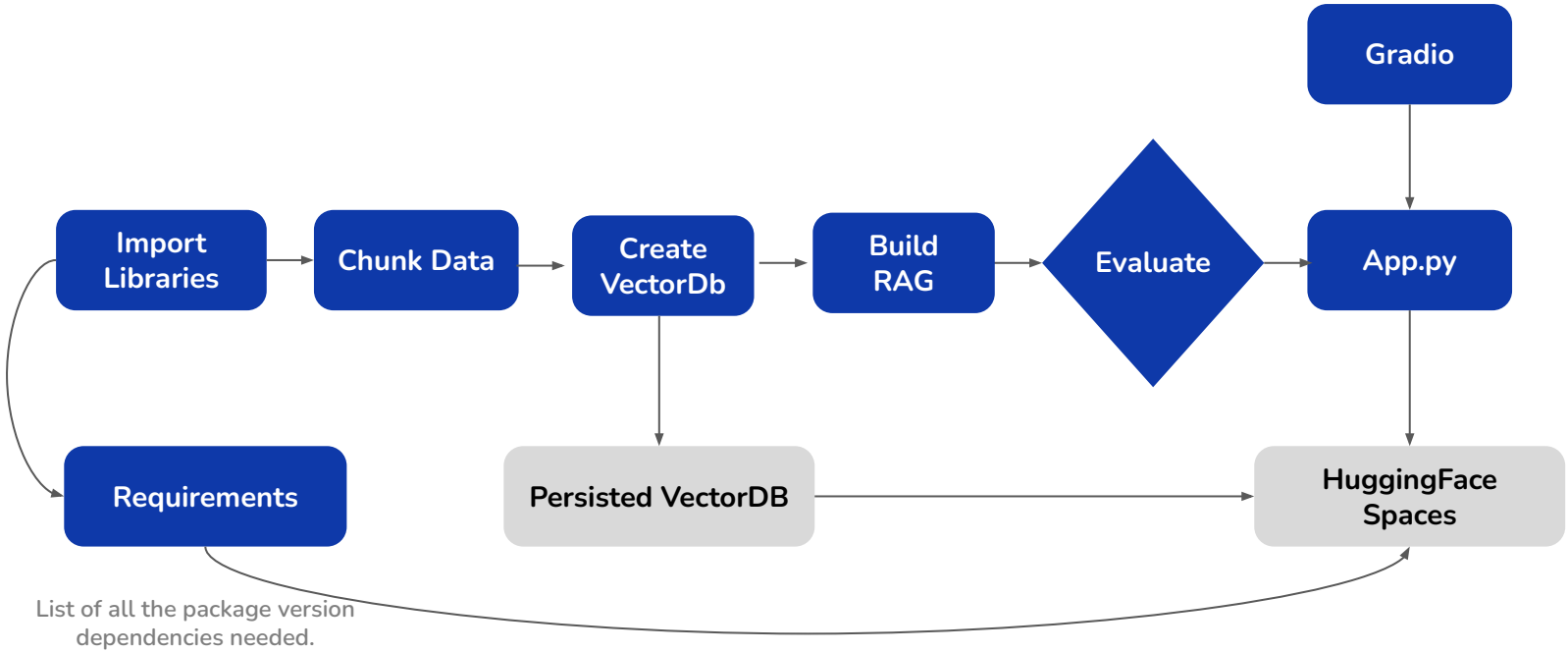
1. Has the company made any significant acquisitions in the AI space, and how are these acquisitions being integrated into the company's strategy?
2. How much capital has been allocated towards AI research and development?
3. What initiatives has the company implemented to address ethical concerns surrounding AI, such as fairness, accountability, and privacy?
4. How does the company plan to differentiate itself in the AI space relative to competitors?

This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Solution Approach

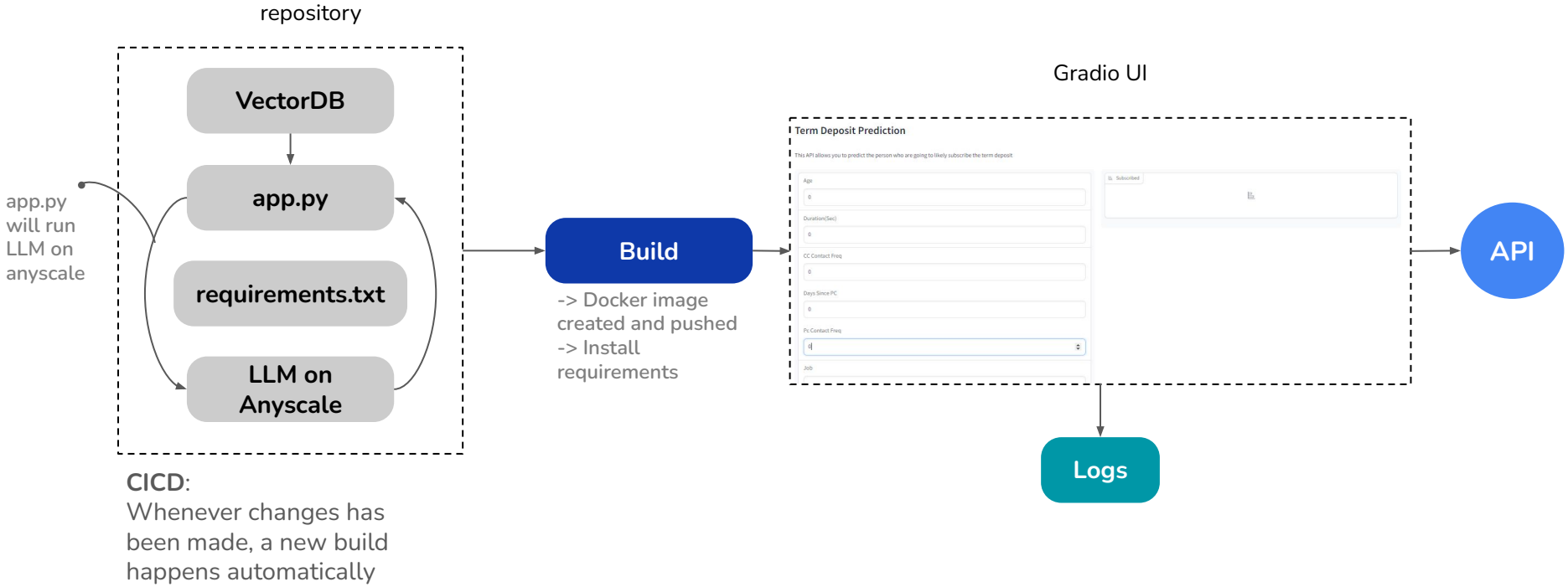
This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Creating and Deploying RAG system



This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

CICD Automation

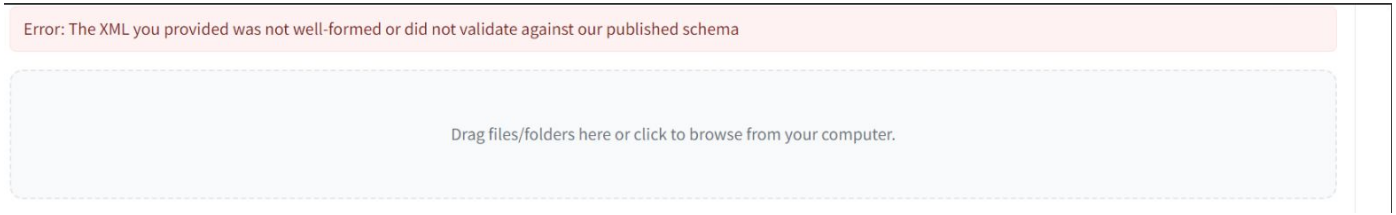


This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Note:

1. If you face the following error while uploading files to Hugging Face, you have to use Hugging Face CLI to upload the files to your spaces as shown below.

Error:



Steps to upload files using Hugging Face CLI:

Install huggingface CLI - `pip install -U "huggingface_hub[cli]"`

Upload files - `huggingface-cli upload repo_name local_path path_in_repo --token=hf_token --repo-type=space`

Example: `huggingface-cli upload username/10kreports ./reports_db ./reports_db/ --token=hf_wPFDRMmQVaflibabflebhtizEFD --repo-type=space`

This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Power Ahead!

This file is meant for personal use by mayank.chugh@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.