Getting Started with RAG

(Retrieval Augmented Generation)

JASON HALEY

BOSTON CODE CAMP
MARCH 2024

Boston Code Camp 36 - Thanks to our Sponsors!

- Platinum

Microsoft

- Gold

MILL5

EQengineered

- Silver

slalom

Progress® Telerik®

PULSAR SECURITY

- In-Kind Donations

MESCIUS

CALL FOR PAPERS POWERED BY sessionize

ChatGPT

ChatGPT officially launched November 30, 2022

By January 2023 had over 100 million users

Why did it take off so fast?

What has changed in the technology environment because of it?

Where do you think its going?

Phases

1. General use and awareness
2. Combining LLM with external data and APIs
3. Agents or multi-step workflows with LLMs and other sources
   - Somewhat autonomous (ideally)
4. Tools that provide actual business value

Demo

SETTING THE STAGE WITH CHAT GPT

Using LLMs in an Application

Bad practice:

- Using LLM as a database or source of facts
  - Can get you into trouble

Better practice

- Have a retrieval step paired with using LLM to reason over a provided context to generate response

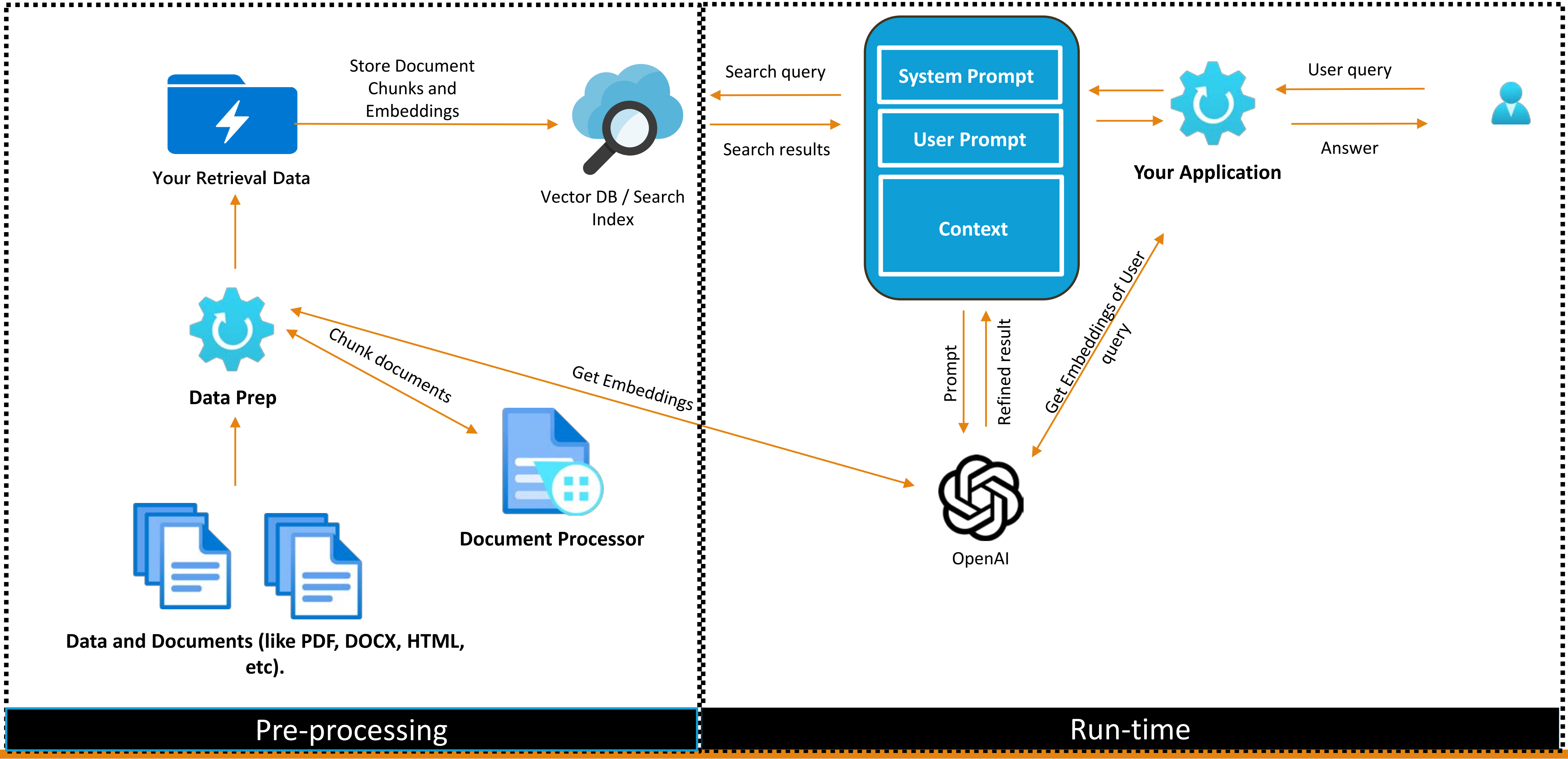
Retrieval

Pre-step of searching your data sources before calling LLM

Goal:

- Provide the best context for LLM

Retrieval Augmented Generation Pattern



Demo:

SIMILARITY SEARCH WITH SQL SERVER

Searching Types

Keyword

Semantic or Similarity

Hybrid (with reranking)

An **embedding** is a sequence of numbers that represent the concepts within a chunk of text

Embeddings and Vector Databases

OPENAI EMBEDDINGS

Text -embedding-ada-002

- 1536 dimensions

Text -embedding-3-small

- 1536 dimensions

Text -embedding-3-large

- 3072 dimensions

VECTOR DATABASE

- Azure AI Search
- Cosmos DB
- Pinecone
- Chroma
- Qdrant
- Many others

Chunking and Its Challenges

Process of breaking something up into smaller parts (ie. files or text)

Keep mind: Embeddings encode a **meaning** for a given chunk of text

What does that mean for blocks of text?

- Character count
- Paragraphs
- Pages
- Overlap
- Other

What about tables of text?

What about images or charts?

Data sources

Documents

- Chunk up files
- Preprocess and index embeddings

Database

- Determine fields the represent the data meaning
- Preprocess and index embeddings

External API

- The provider will be responsible for doing one of the above

Demo

SEARCHING DOCUMENTS IN RETRIEVAL

Demo

SEARCHING DATA OBJECTS IN RETRIEVAL

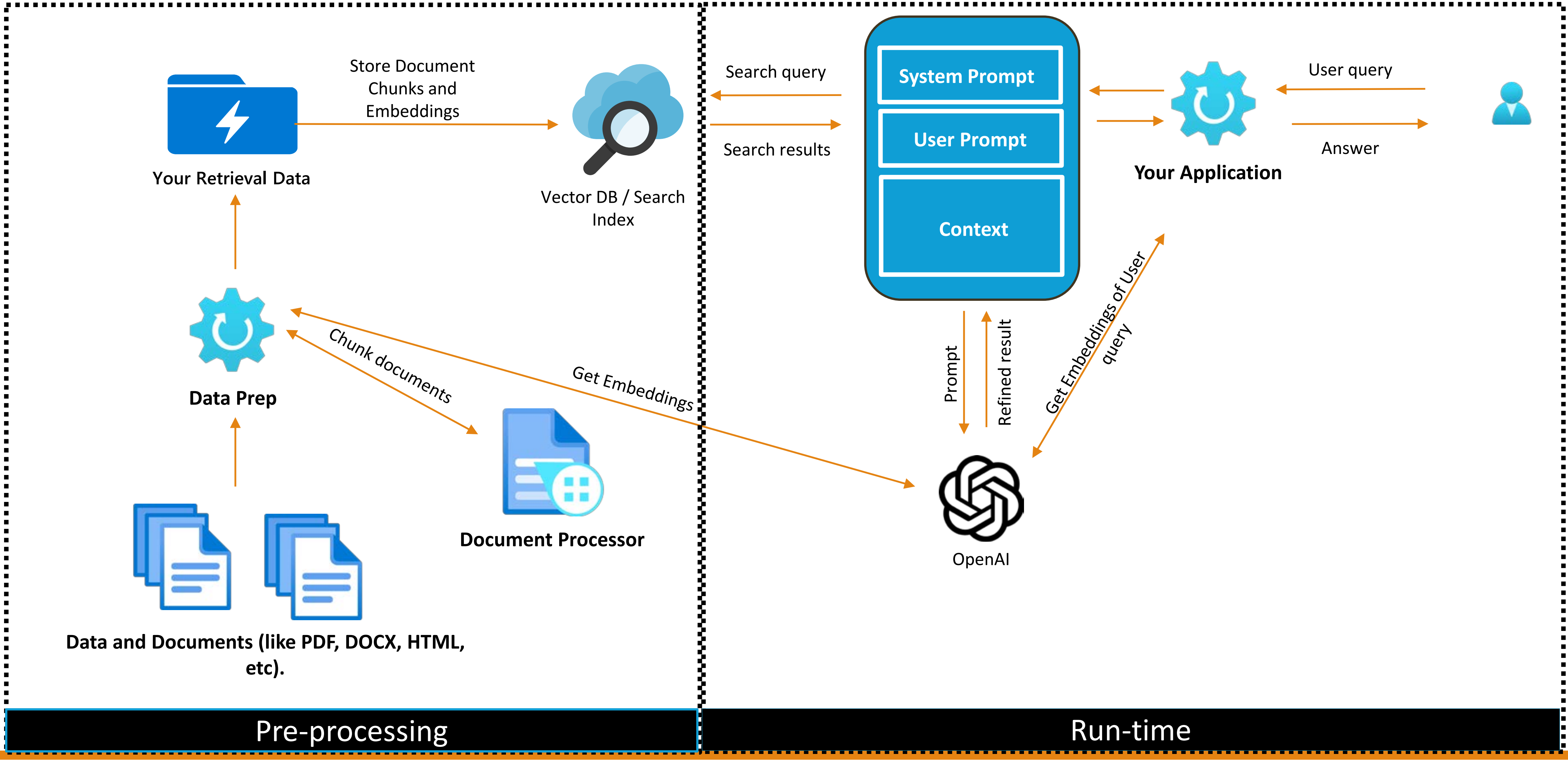
Generation

Given a relevant context to the user question, generate a coherent and useful response

Goal:

Provide the full capability of the LLM using the domain specific context provided from the retrieval step

Retrieval Augmented Generation Pattern



Chat Completion Request

Messages

- **System** – instructions and context for the LLM
- **User** – user query
- **Assistant** – LLM response

Other API Parameters

- **Max Tokens** – for the chat completion to use
- **Response Format** (text or json_object)
  - NOTE: you must also instruct the model to produce JSON in the system or user message or it will return an error
- **Stream** – will send message deltas
  - Good to make application feel faster, though complicates the implementation
- **Temperature** – between 0 – 2
  - Higher the value the more randomness
- **Top P** –Top probability 0 – 1
  - Alternate to using temperature
- **Tools** – list of tools for the LLM to call back to
  - Topic for another talk

Other Application Concerns

Loading/Saving History/Feedback

- Database backed

Token Usage

Post-processing

- May want to show the user only part of the result

More Application Concerns

Data Privacy

Latency

LLMOps

Evaluation

- Groundedness – how well a model’s generated answers align with information from context given
- Relevance – the extent to which the model’s answers are pertinent and related to the question
- Coherence – how well the language model’s answers read naturally and is clearly understood
- Fluency – is the language proficiency of a model’s answers
- Similarity – quantifies the similarity between the ground truth and the model’s answer

Resources

Azure OpenAI RAG Pattern using a SQL Vector Database

- https://blazorhelpwebsite.com/ViewBlogPost/10067

azure-search-openai-demo-csharp

- https://github.com/Azure-Samples/azure-search-openai-demo-csharp

Vector -Search-AI-Assistant

- https://github.com/Azure/Vector-Search-AI-Assistant

The AI Chat App Hack

- https://github.com/microsoft/AI-Chat-App-Hack#hack-together-the-ai-chat-app-hack