Simplifying RAG with Databricks Vector Search

Wilson Mok

Mar 14, 2025

Dear Azure

Bio

Wilson Mok

Sr. Data architect & Consultant

in /wilson-mok

/wilson-mok

@the-analytics-lab

MVP
Microsoft®
Most Valuable Professional

My experiences includes:

- Avanade
- CAE
- Air Canada

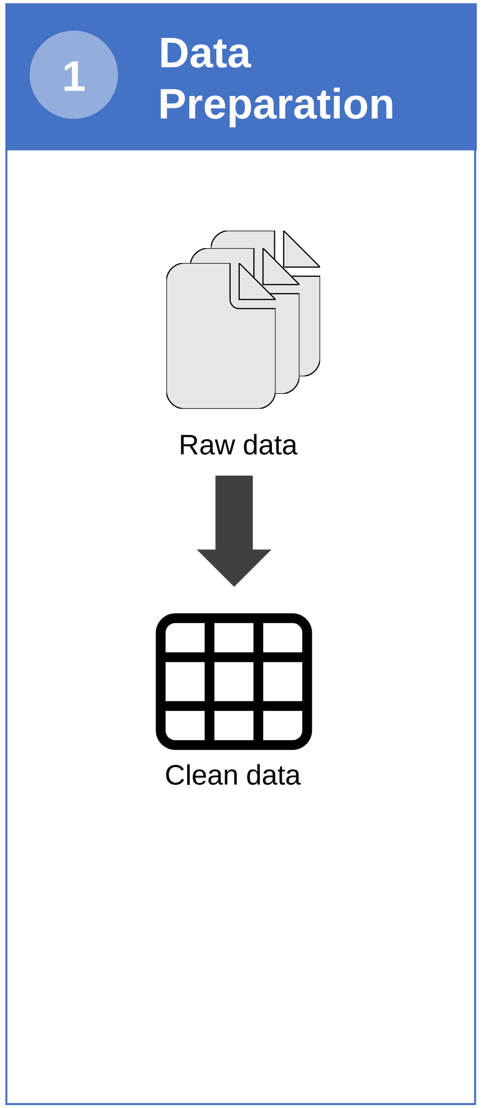
databricks
PARTNER PROGRAM
SOLUTIONS ARCHITECT CHAMPION

Expected Audience and Presentation Structure

- This session, we will cover:
  - Discuss the complexity of RAG pipelines.
  - Simplifying the RAG pipelines using Databricks Vector Search.
  - Using Unity Catalog’s delta table and volume to store and manage our data.
  - Quick Introduction on Databricks Genie.
  - Demo: Implementing RAG using Databricks Unity Catalog and Vector Search.

- Fundamental knowledge: Gen AI, Databricks and Python would be beneficial.

There are four key steps in a RAG workflow to create an AI Chatbot.

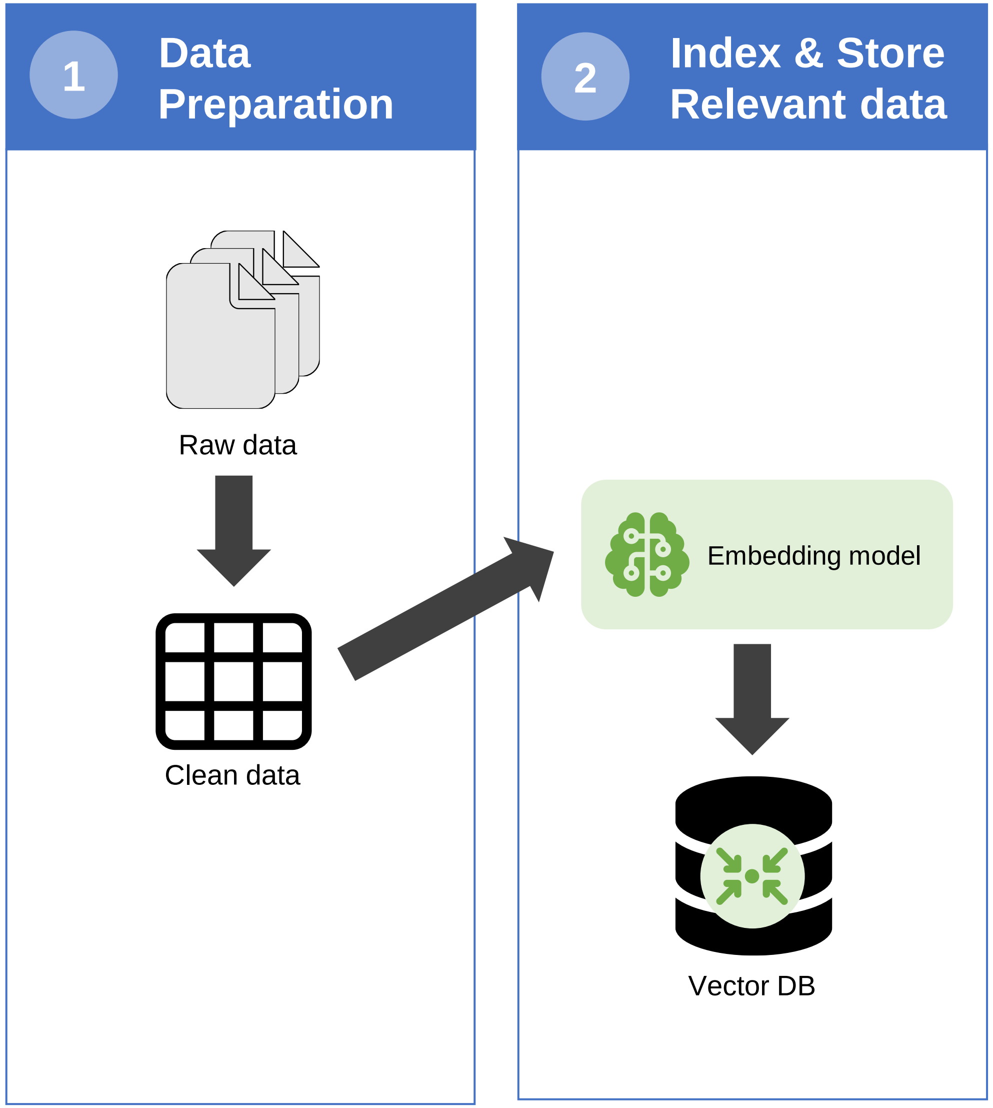


1 Data Preparation

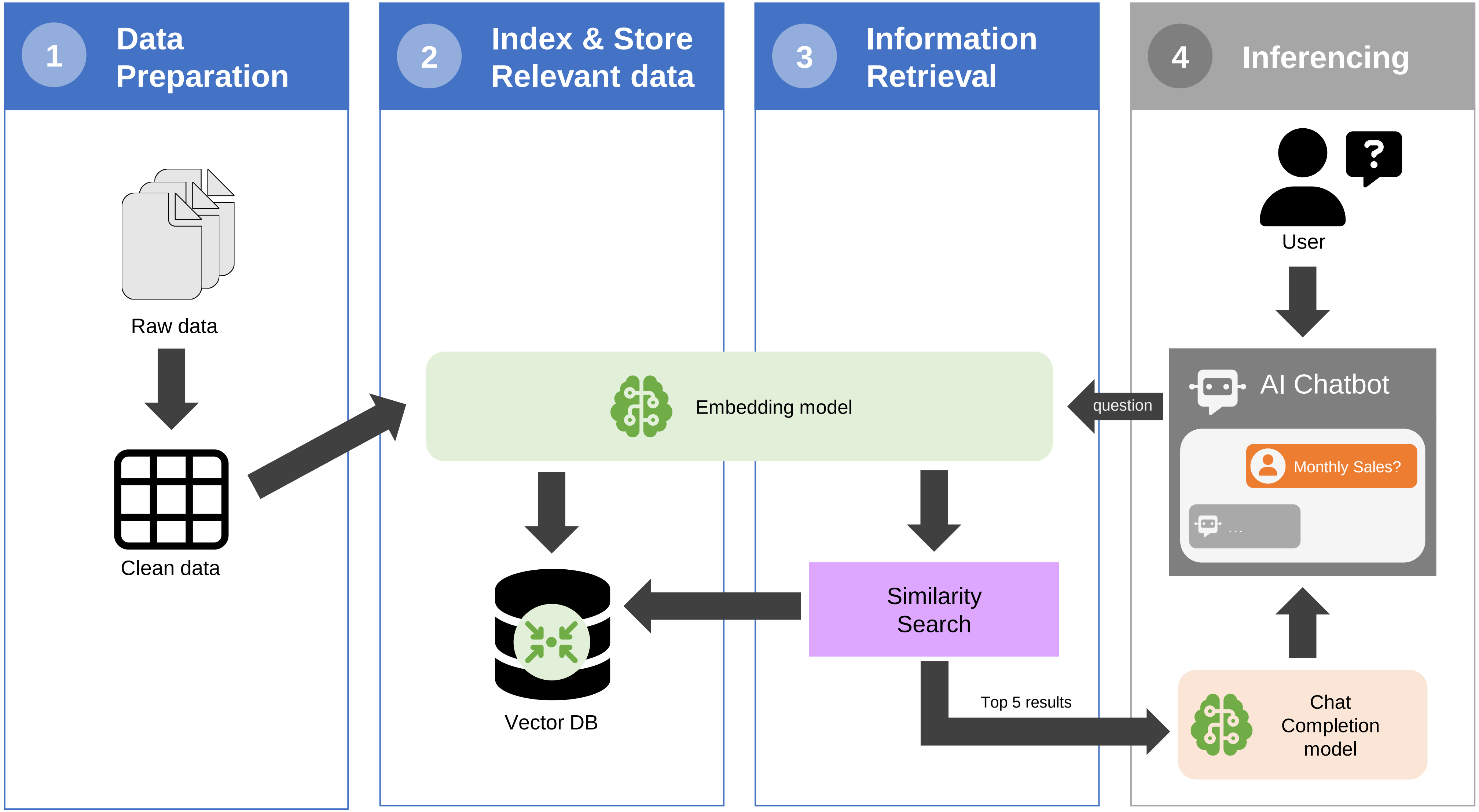
Raw data

Clean data

There are four key steps in a RAG workflow to create an AI Chatbot.



There are four key steps in a RAG workflow to create an AI Chatbot.

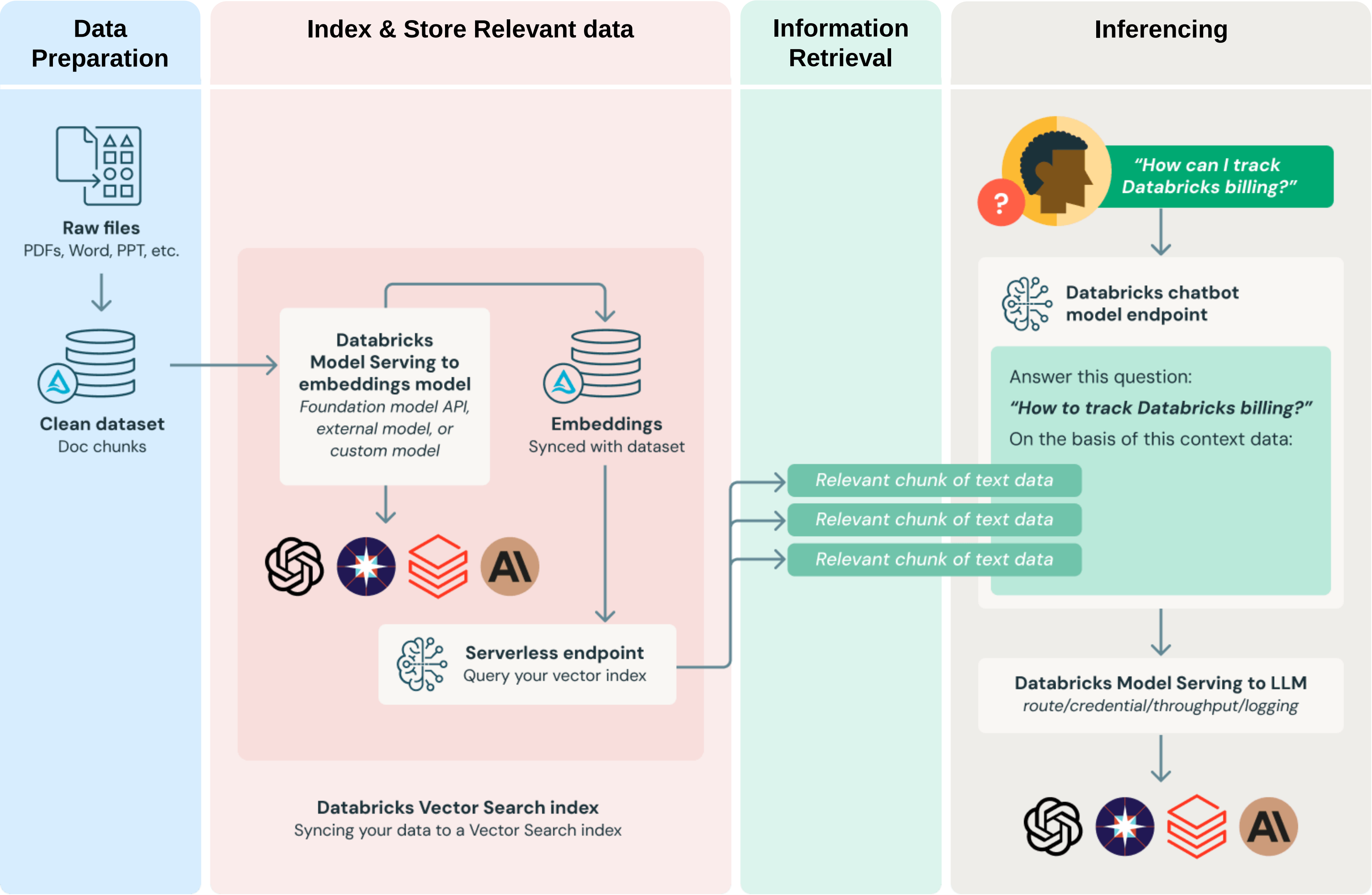


* This presentation will not be covering Chat Completion in detail

Each step in the RAG workflow requires a set of specialized tools and technologies…

Data processing	• Azure Cognitive Search • Azure AI Vision • Azure AI Speech-to-Text • Spark (Databricks, MS Fabric) • Orchestrator (Azure Data Factory, LangChain)
Vector Database	• Pinecone • Azure AI Search • Postgres DB (pgvector)
Embedding Model	• Open AI: text-embedding-3 • Google: text-embedding-gecko • Open Source: GTE
LLM Model	• Open AI: GPT-4o • Google: Gemini • Anthropic: Claude • Open Source: Meta Llama 3.3, DeepSeek

Databricks streamlines this workflow by…



Data Preparation

Raw files
PDFs, Word, PPT, etc.

Clean dataset
Doc chunks

Index & Store Relevant data

Databricks Model Serving to embeddings model
Foundation model API, external model, or custom model

Embeddings
Synced with dataset

Serverless endpoint
Query your vector index

Databricks Vector Search index
Syncing your data to a Vector Search index

Information Retrieval

Relevant chunk of text data

Relevant chunk of text data

Relevant chunk of text data

Inferencing

"How can I track Databricks billing?"

Databricks chatbot model endpoint

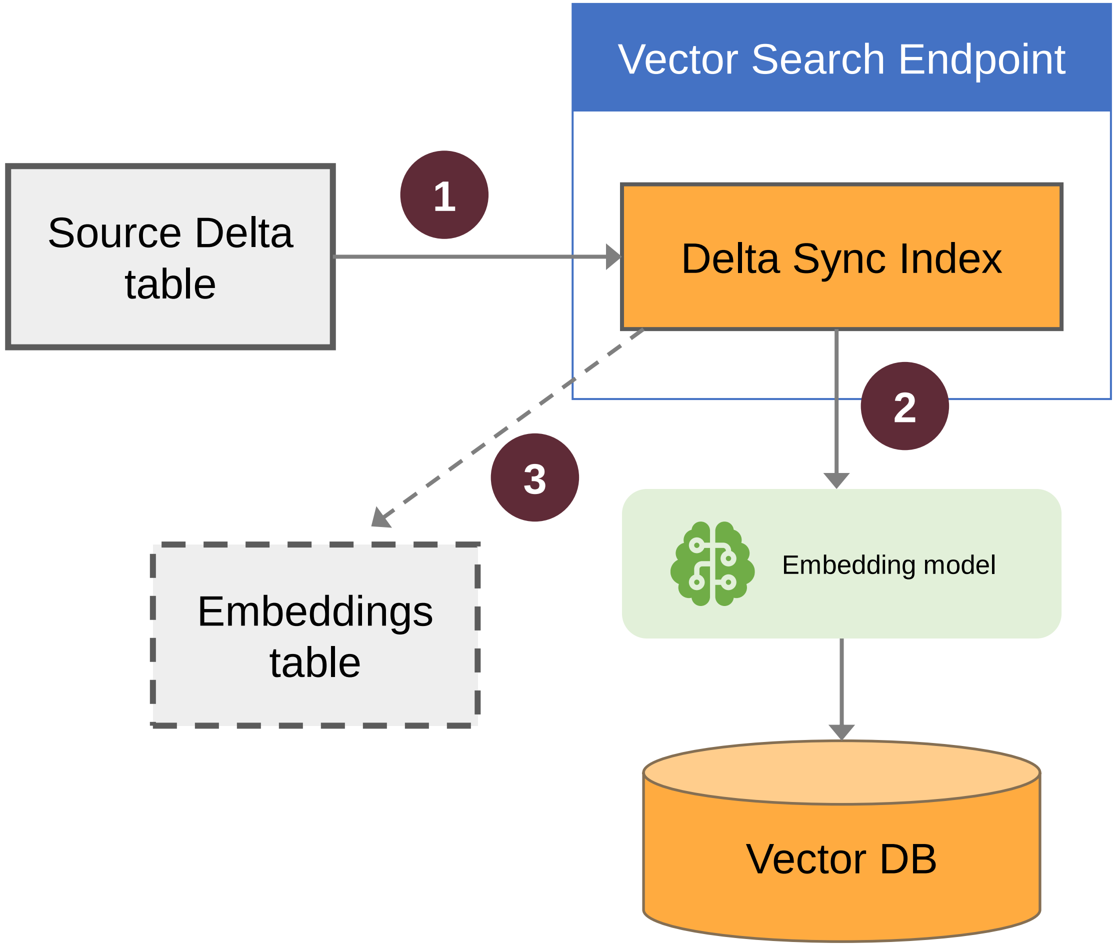
Answer this question:
"How to track Databricks billing?"
On the basis of this context data:

Databricks Model Serving to LLM
route/credential/throughput/logging

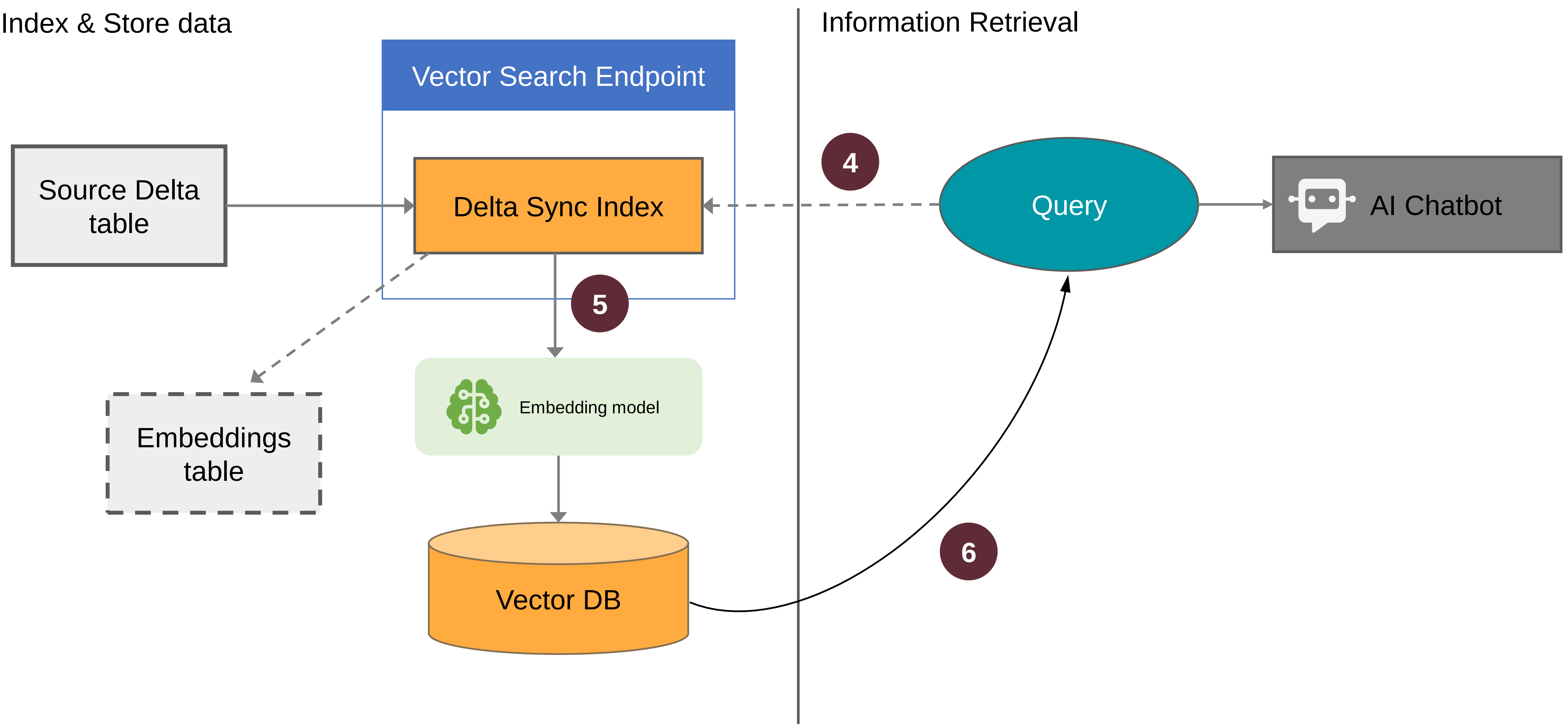
Credit: https://www.databricks.com/glossary/retrieval-augmented-generation-rag

A Closer look into Databricks – Mosaic AI Vector Search…

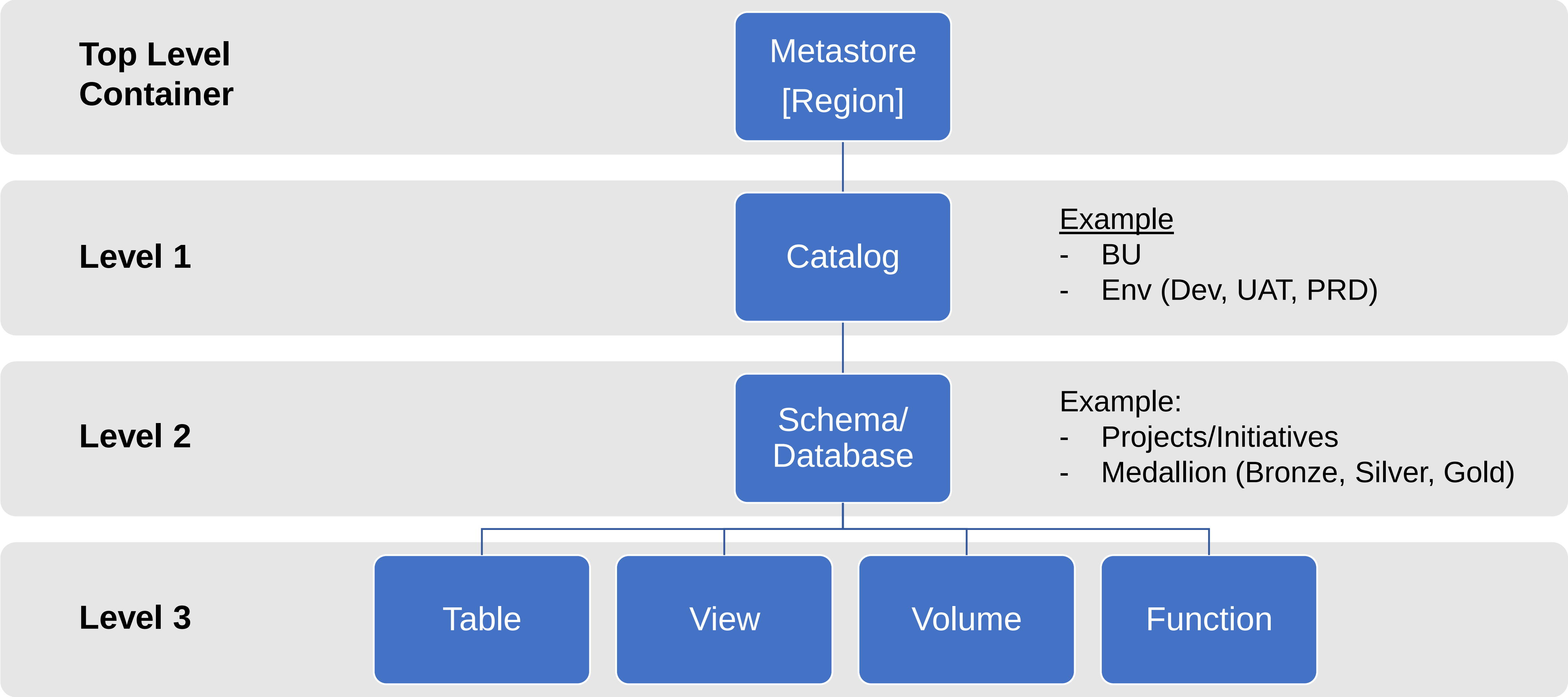
Index & Store data



A Closer look into Databricks – Mosaic AI Vector Search…



Unity Catalog provides a way to govern and manage…



Top Level Container

Metastore [Region]

Level 1

Catalog

<u>Example</u>
- BU
- Env (Dev, UAT, PRD)

Level 2

Schema/ Database

Example:
- Projects/Initiatives
- Medallion (Bronze, Silver, Gold)

Level 3

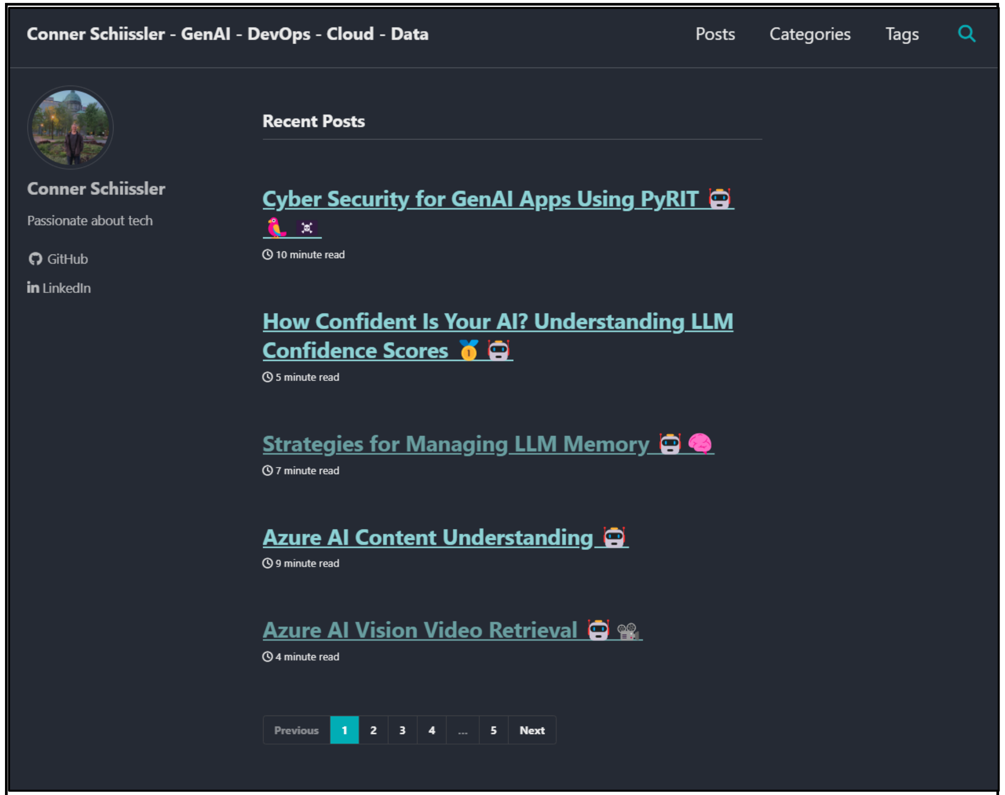
Table

View

Volume

Function

Demo 1: Implementing RAG using Databricks Unity Catalog and Vector Search



Conner Schiissler - GenAI - DevOps - Cloud - Data

Posts Categories Tags

Conner Schiissler

Passionate about tech

GitHub

LinkedIn

Recent Posts

Cyber Security for GenAI Apps Using PyRIT 🤖 🦜 ☠

10 minute read

How Confident Is Your AI? Understanding LLM Confidence Scores 🥇 🤖

5 minute read

Strategies for Managing LLM Memory 🤖 🧠

7 minute read

Azure AI Content Understanding 🤖

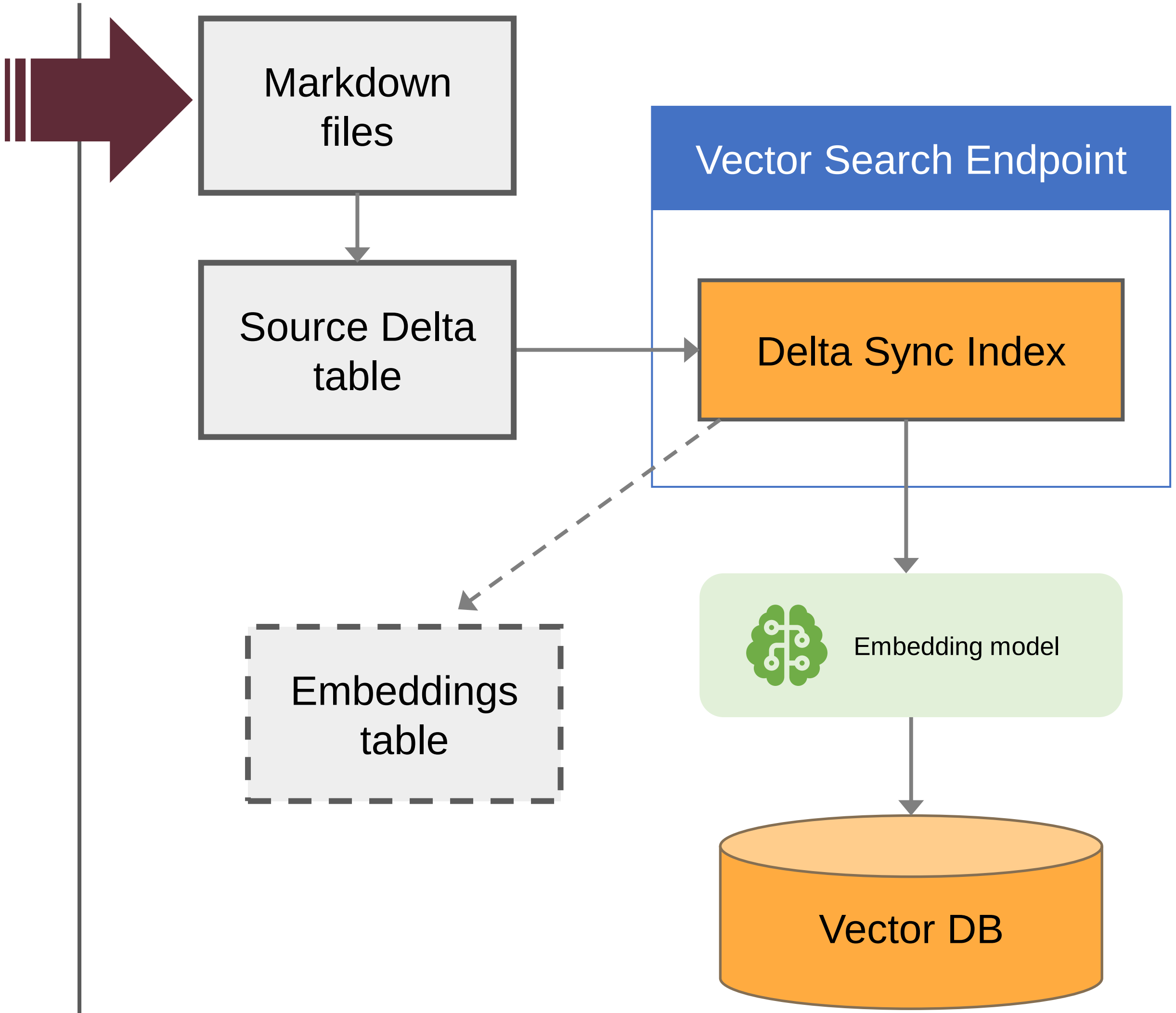
9 minute read

Azure AI Vision Video Retrieval 🤖 📽

4 minute read

Previous 1 2 3 4 ... 5 Next

https://schiiss.github.io/blog/



Q & A 1

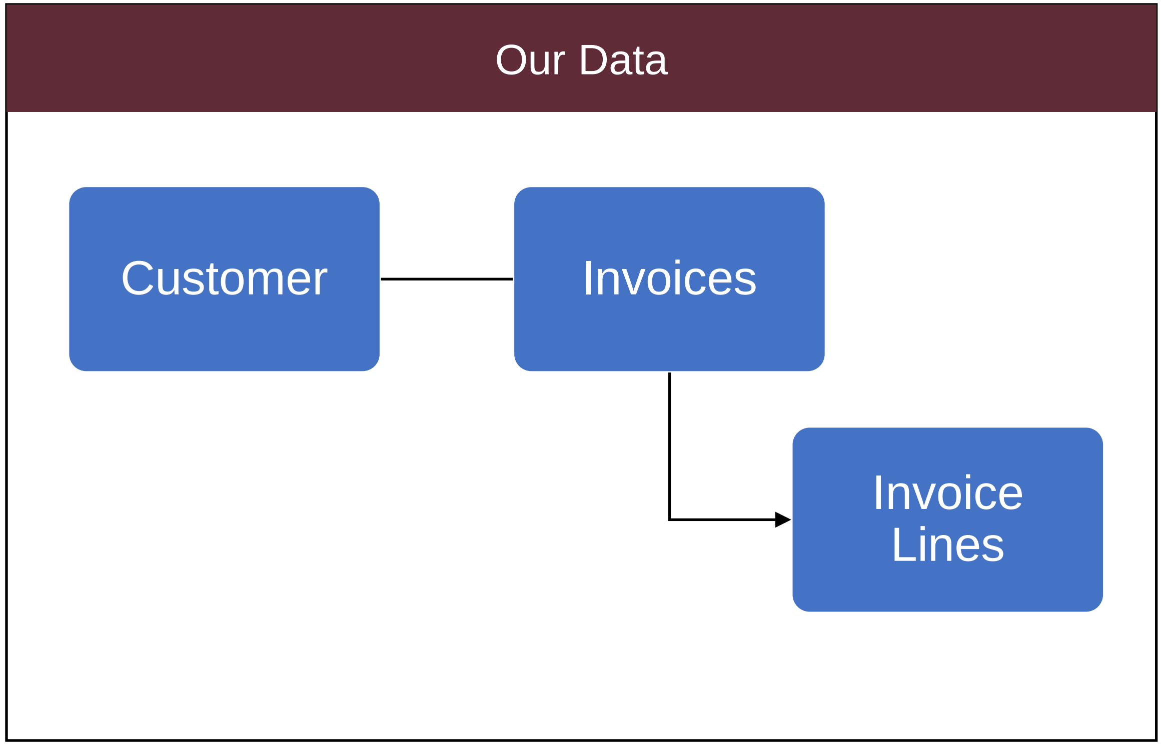
Databricks: AI/BI Genie

Credit: https://www.databricks.com/blog/onboarding-your-new-aibi-genie

What is an AI/BI Genie?

- Conversational AI assistant powered by Open AI.
- Talk to your data in Natural Language.
- Genie translates them into SQL and retrieve relevant insights.
- Can generate Visualizations like Charts and Graphs.

Demo: Databricks’ AI/BI Genie



Credit: https://www.databricks.com/blog/onboarding-your-new-aibi-genie

Q & A 2

After all this, we have run through…

1. Mosaic AI Vector Search: Starts with a Delta Table, creates a Delta Sync Index to create embedding via Vector Search Endpoint.
2. Unity Catalog: Provides a governed, structured approach to manage your data using three-level hierarchy. We used Delta tables and Volumes in action today.
3. AI/BI Genie: An AI-powered assistant that enables natural language querying, automated insights, and seamless BI integration.

- The code is in my Github repo.
  - https://github.com/wilson-mok/demo/tree/main/2025/DearAzure/1-Databricks-Vector-Search

Thank you

LinkedIn