

Building a RAG Chatbot with Databricks

Vika Koval

Проблема та Мета

Проблема: LLM мають лише публічно доступні знання з певним часовим обмеженням.

Обмеження: Ми не можемо згодувати моделі тисячі документів за один раз та публічно поширювати корпоративну інформацію.

Рішення:

Використання RAG

RAG

Retrieval-Augmented Generation (RAG)

RAG об'єднує:

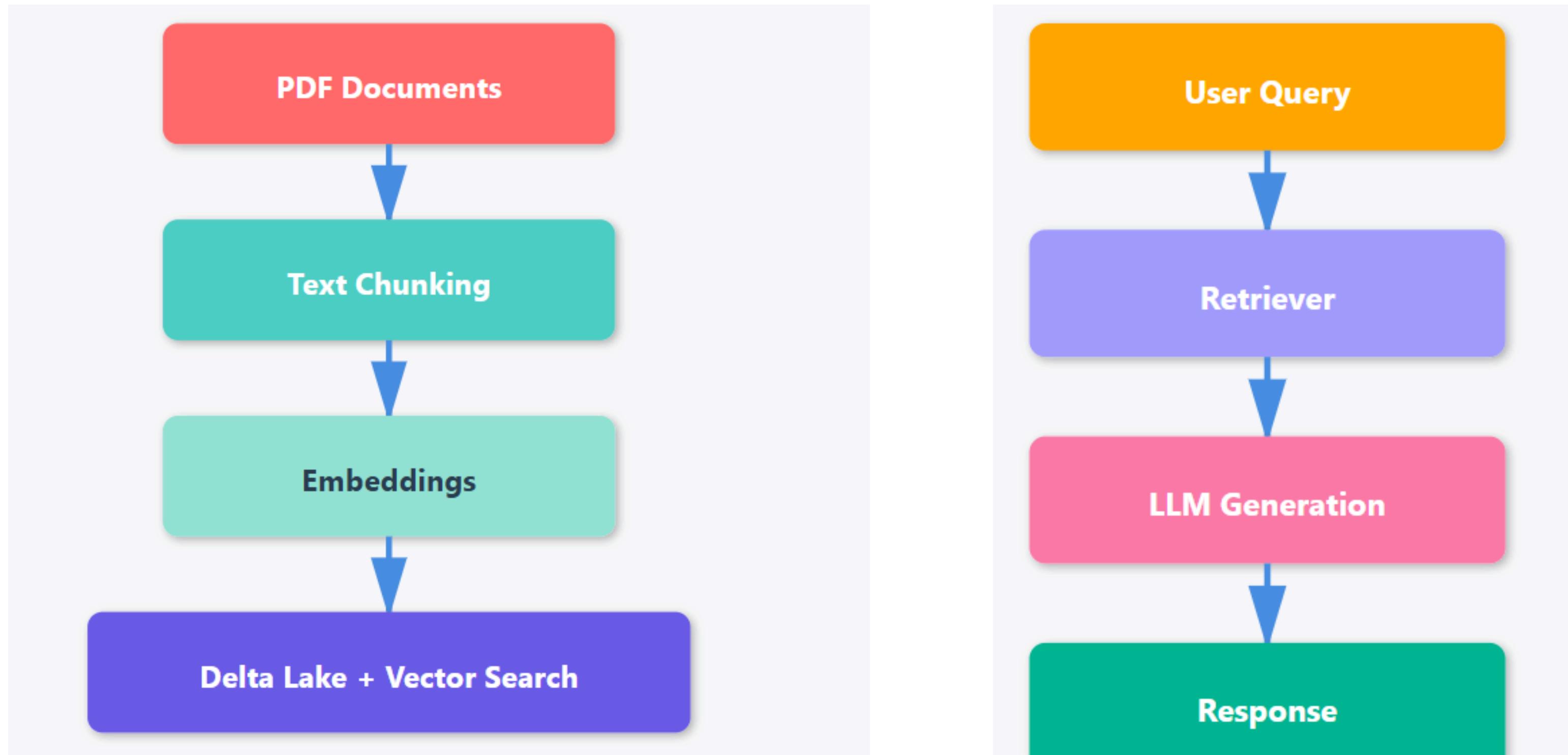
Retrieval - пошук релевантної інформації з бази знань

Generation - генерація відповіді через LLM на основі знайденої інформації

Переваги:

1. Актуальна інформація
2. Менші галюцинації моделі
3. Джерела інформації можна верифікувати
4. Не потрібно перенавчати модель

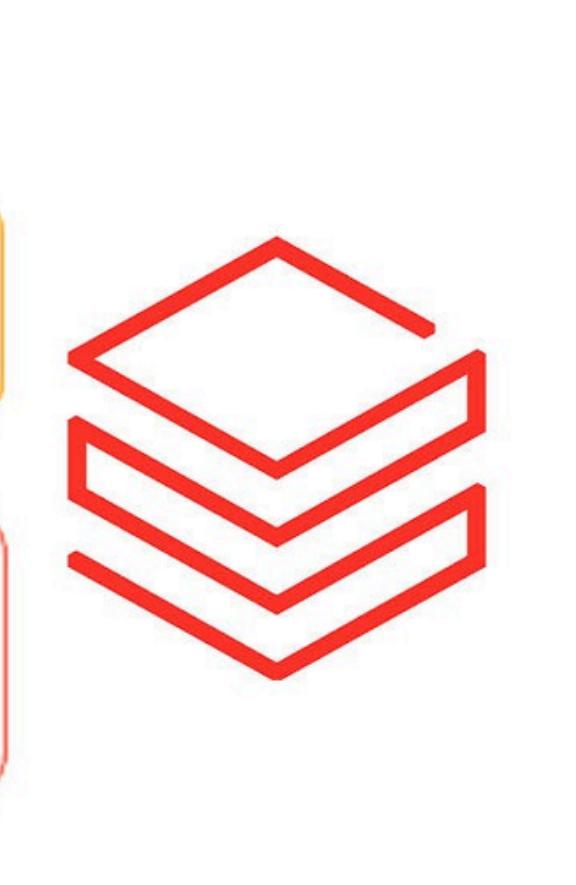
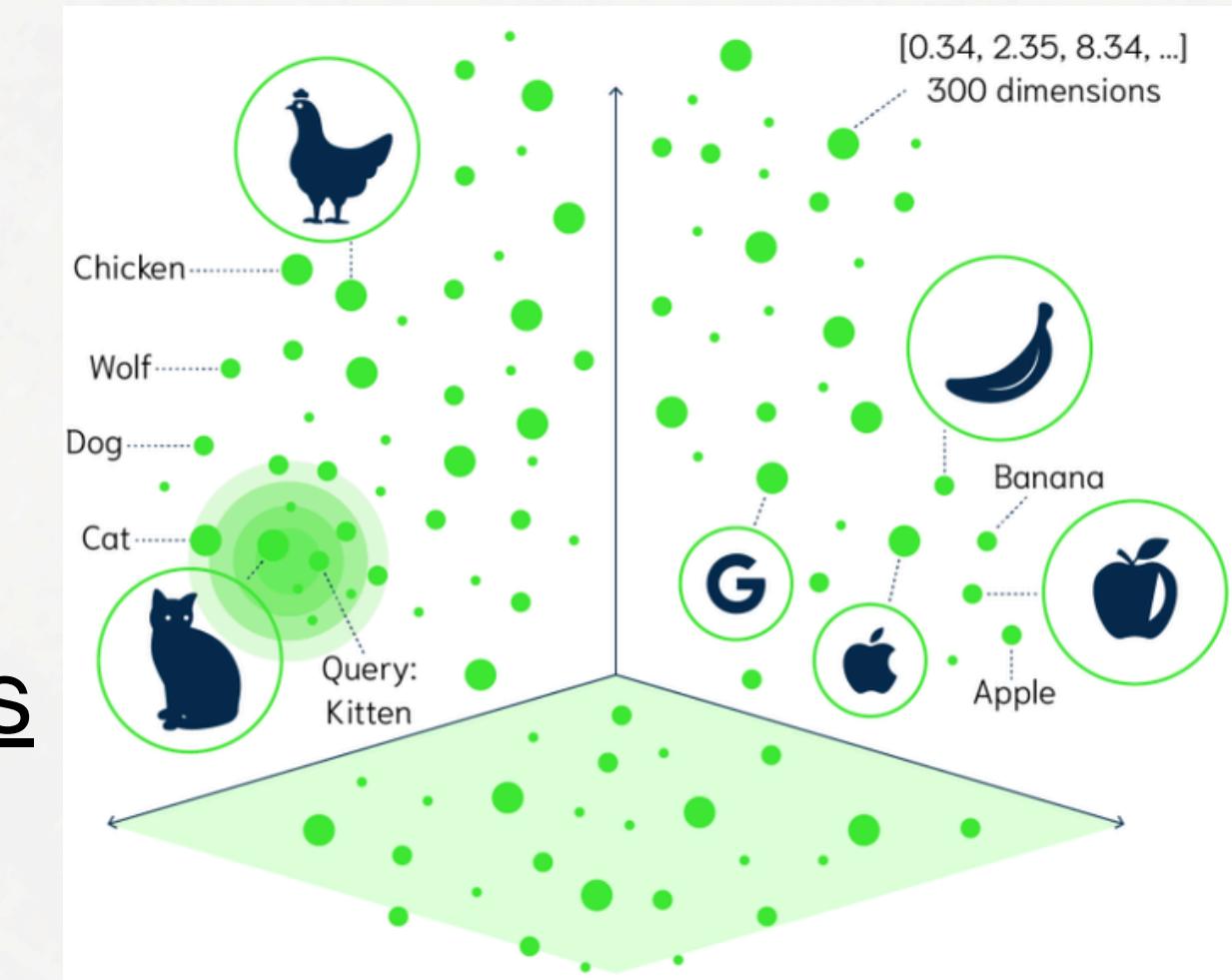
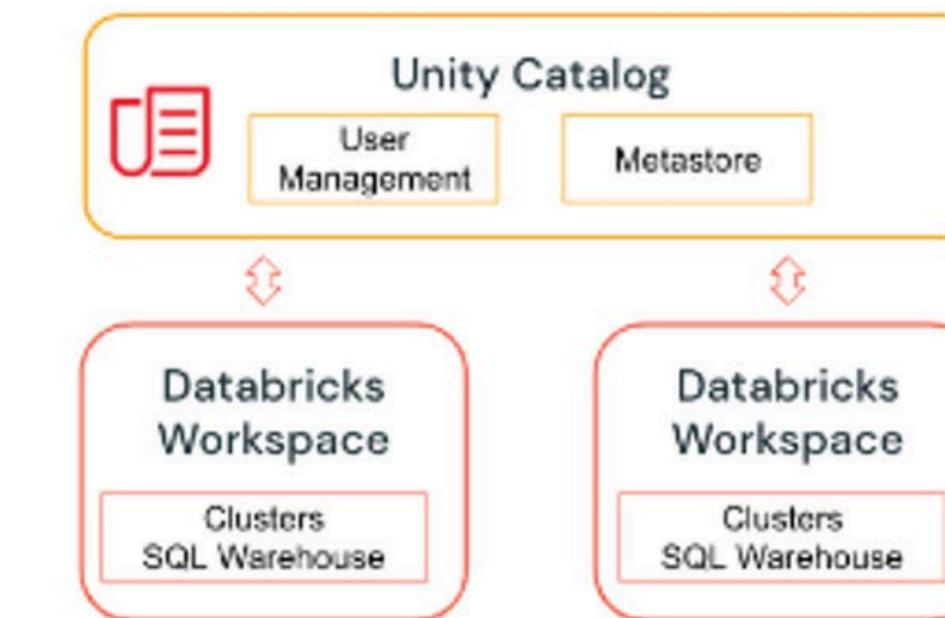
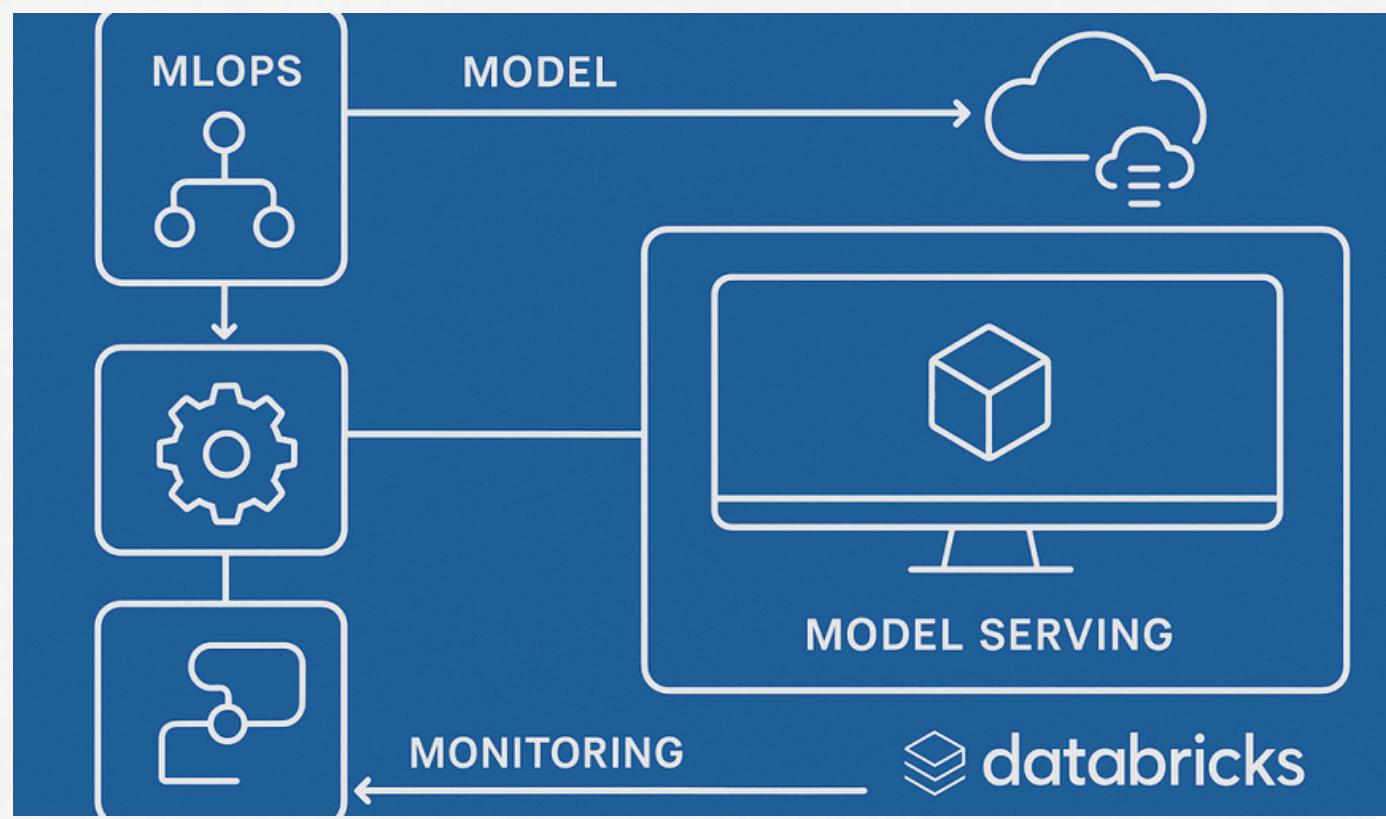
Архітектура



Дані

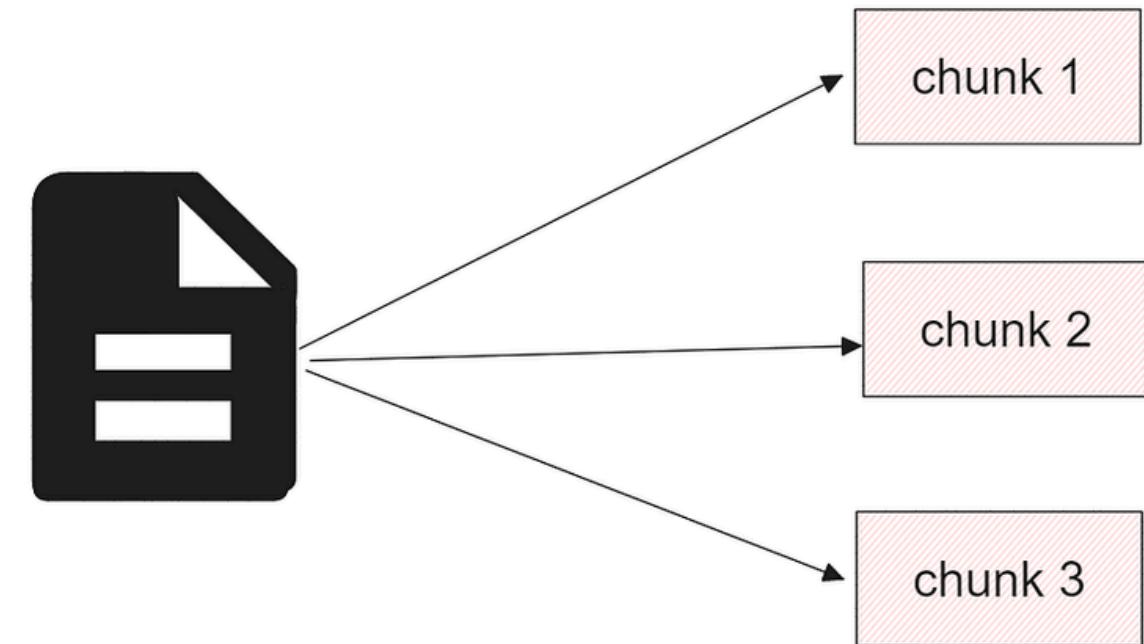
Документація Databricks:

- Unity Catalog : [What is Unity Catalog?](#)
- Vector Search : [Databricks Vector Search](#)
- Model Serving: [Model Serving with Databricks](#)



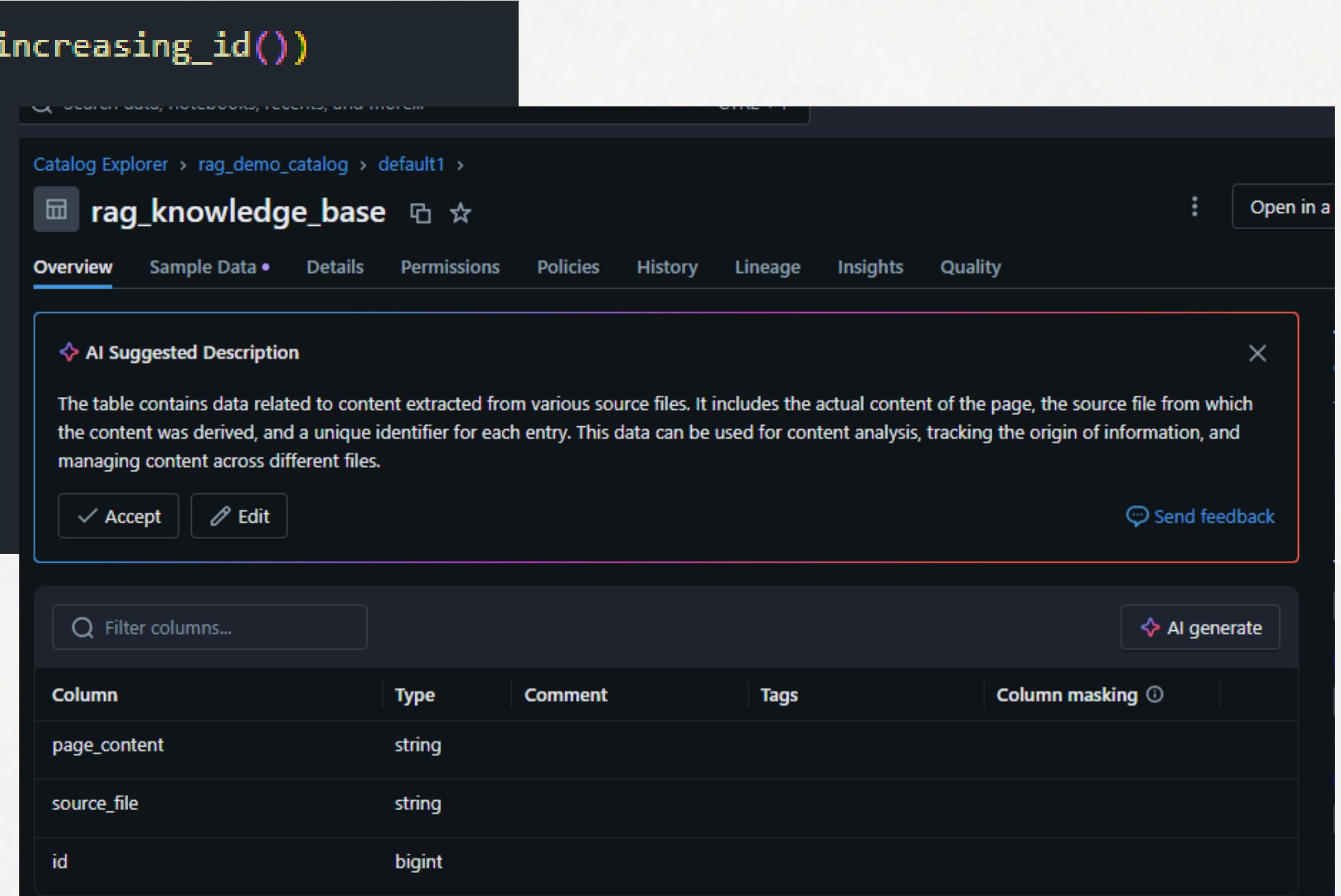
Підготовка Даних

```
text_splitter = RecursiveCharacterTextSplitter(  
    chunk_size=1000,  
    chunk_overlap=200,  
    separators=[ "\n\n", "\n", " ", "" ]  
)  
  
for file_path in files:  
    loader = PyPDFLoader(file_path)  
    docs = loader.load()  
    chunks = text_splitter.split_documents(docs)
```



Vector Database

```
df_with_id = df.withColumn("id", F.monotonically_increasing_id())  
  
(df_with_id.write  
    .format("delta")  
    .mode("overwrite")  
    .option("overwriteSchema", "true")  
    .option("delta.enableChangeDataFeed", "true")  
    .saveAsTable(table_name)  
)
```



The screenshot shows the Databricks Catalog Explorer interface. The top navigation bar includes 'Catalog Explorer', 'rag_demo_catalog', 'default1', and a search bar. The main title is 'rag_knowledge_base'. Below the title is a toolbar with 'Overview' (selected), 'Sample Data', 'Details', 'Permissions', 'Policies', 'History', 'Lineage', 'Insights', and 'Quality'. An 'Open in a' button is also present. A modal window titled 'AI Suggested Description' is open, containing the following text: 'The table contains data related to content extracted from various source files. It includes the actual content of the page, the source file from which the content was derived, and a unique identifier for each entry. This data can be used for content analysis, tracking the origin of information, and managing content across different files.' There are 'Accept' and 'Edit' buttons at the bottom of the modal. In the main table area, there is a header row with columns: 'Column', 'Type', 'Comment', 'Tags', and 'Column masking'. Below the header, three rows of data are listed: 'page_content' (string type), 'source_file' (string type), and 'id' (bigint type). A 'Filter columns...' search bar is located above the table, and an 'AI generate' button is in the top right corner of the table area.

Column	Type	Comment	Tags	Column masking
page_content	string			
source_file	string			
id	bigint			

Зберігаємо дані у Delta Table.

CDF увімкнено для автоматичної синхронізації з індексом

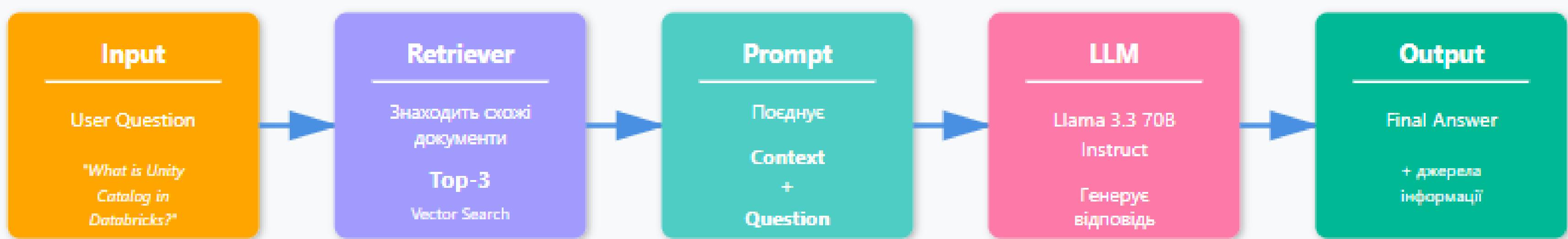
Vector Search

```
vsc.create_delta_sync_index(  
    endpoint_name=ENDPOINT_NAME,  
    source_table_name=SOURCE_TABLE,  
    index_name=INDEX_NAME,  
    pipeline_type="TRIGGERED",  
    primary_key="id",  
    embedding_source_column="page_content",  
    embedding_model_endpoint_name="databricks-bge-large-en")
```



RAG Chain

```
chain = (
    {"context": retriever | format_docs, "question": RunnablePassthrough()}
    | prompt_template
    | llm
    | StrOutputParser()
)
```



Деплой

Frontend: Streamlit

Хостинг: Databricks Apps

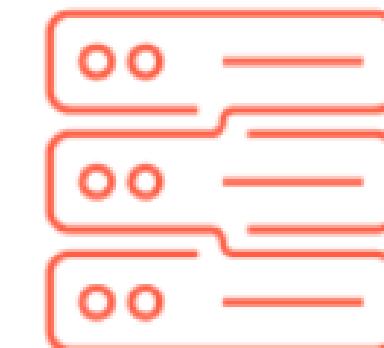
Безпека: Додаток працює в ізольованому контейнері

Авторизація: Використання Service Principal для доступу до Vector Search

Інтеграція: Прямий доступ до Unity Catalog



Streamlit



Databricks
Apps



Виклики та Рішення

1. Конфлікт портів:

Проблема: App Not Available.

Рішення: Налаштування динамічного порту через зміну \${DATABRICKS_APP_PORT} у app.yaml.

2. Авторизація у Vector Search:

Проблема: Додаток не бачив контексту Workspace.

Рішення: Явна передача DATABRICKS_HOST та TOKEN через environment variables у команду запуску.

3. YAML Formatting:

Проблема: Помилки парсингу конфігурації.

Рішення: Генерація app.yaml скриптом для уникнення помилок з проблами.

DEMO

RAG Chatbot



for what I need to use unity catalog



You need to use Unity Catalog for the following purposes:

1. **Centralized data governance:** Unity Catalog provides a centralized data catalog that offers access control, auditing, lineage, quality monitoring, and data discovery capabilities across Databricks workspaces.
2. **Access control and security:** Unity Catalog allows you to define data access policies that apply across all workspaces in a region, using a standards-compliant security model based on ANSI SQL.
3. **Data discovery and management:** Unity Catalog provides a single place to administer data access policies, making it easier to manage and discover data across your Databricks workspaces.
4. **AI asset management:** Unity Catalog also manages AI assets, such as models and functions, in addition to data assets like tables and views.

Overall, Unity Catalog helps you to govern your data and AI assets in a unified and centralized way, making it easier to manage access, security, and discovery across your Databricks workspaces.

Q&A

THANK YOU FOR LISTENING