Transforming Data Warehouses into Dynamic Knowledge Bases for RAG

Gerry Hosea, Student, IT, University of North Sumatra, Medan, North Sumatra

Hari Sudrajat, Software Developer, IT, TRT Solution Limited

Abstract

It is necessary to include data warehouses in contemporary data processing frameworks to provide comprehensive support for efficient decision-making procedures. This study aims to evaluate the exploitation of data warehouses as a knowledge base inside a Retrieval-Augmented Generation (RAG) model. This model combines retrieval mechanisms with generative models to improve information retrieval and response generation in artificial intelligence systems. Several necessary procedures are the subject of this research. These procedures include the preparation of data via the use of Databricks, the production of online tables, and the transformation of these tables into embeddings. Databricks offers a robust data engineering platform, enabling practical data input, cleaning, and structuring into Delta tables. This is followed by building online tables, making it easier to get data quickly. They are then converted into embeddings, which can capture the semantic substance of the data. These online tables are subsequently altered. The embeddings are kept in a repository, and the RAG model makes use of them to create replies consistent with the context in which they are being used. RAG models can efficiently harness enormous data repositories, as shown by the findings of this research, which reveal considerable increases in the speed at which data is retrieved and the precision of responses. By implementing best practices and using Databricks' capabilities, businesses can improve their AI-driven decision-making processes. This approach is advantageous for various purposes, including customer assistance, data analysis, and strategic planning. By doing more study in the future, it will be possible to investigate the applicability of this technique across a variety of domains and the incorporation of sophisticated generative models to enhance performance.

Keywords: *Data Warehouses, Knowledge Base, Retrieval-Augmented Generation, Databricks, Embeddings, AI Response Generation*

Introduction

There is a once-in-a-lifetime chance to improve information retrieval and decision-making in big data and AI by integrating data warehouses with knowledge bases. By integrating retrieval techniques with neural networks, Retrieval-Augmented Generation (RAG) models have become a potent tool for using massive datasets. Within an RAG framework, this article investigates data warehouses as a knowledge base and the approaches and advantages of doing so.

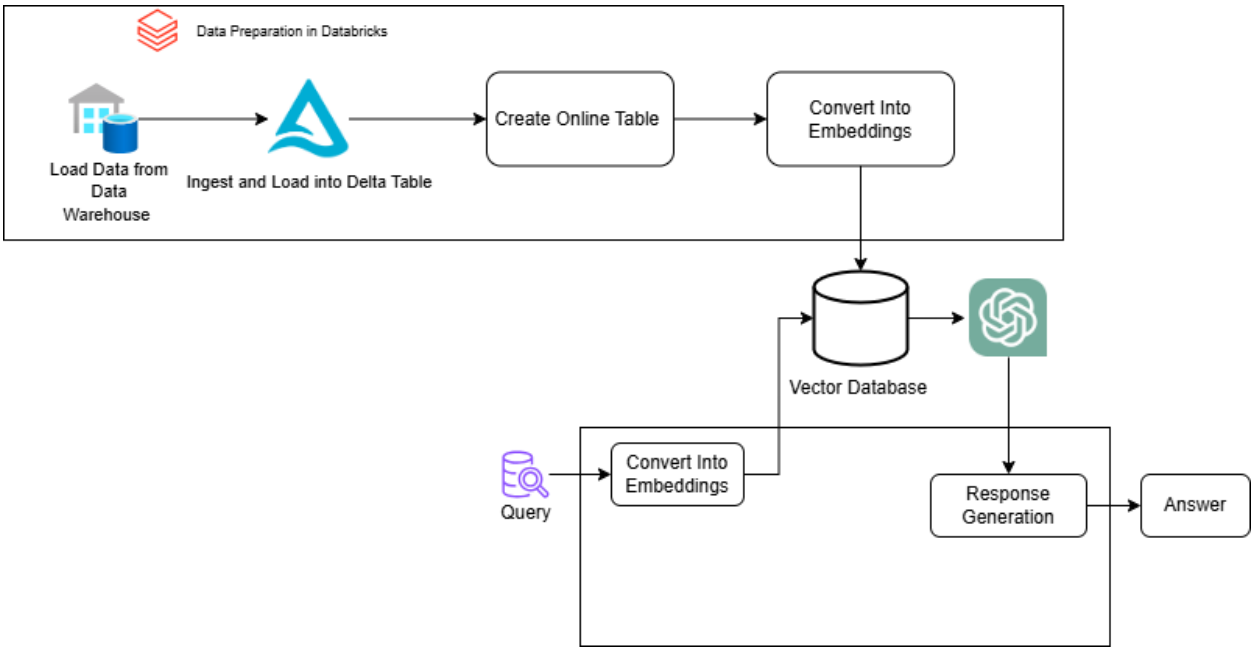

Image 1. Data warehouses with RAG

The image above illustrates handling data warehouses as a knowledge base in a Retrieval-Augmented Generation (RAG) model. The workflow involves several key steps:

1. **Data Loading from Data Warehouse**: The process begins by loading data from a data warehouse into the Databricks environment.
2. **Ingest and Load into Delta Table**: The data is ingested and loaded into a Delta table, ensuring it is structured and clean.
3. **Create Online Table**: An online table from the Delta table facilitates quick and efficient data retrieval.
4. **Convert into Embeddings**: The online table is converted into embeddings, capturing the semantic meaning of the data.
5. **Embedding Storage and Response Generation**: The embeddings are stored in a repository and are used by the RAG model to generate accurate and contextually relevant responses.

This workflow highlights the importance of structured data preparation and embedding conversion in enhancing the performance of RAG models.

Data Preparation in Databricks

Databricks offers a comprehensive platform for data engineering, enabling efficient data preparation and integration processes. The data preparation pipeline involves

loading data from a data warehouse, ingesting and loading it into a Delta table, and creating an online table. This structured approach ensures that the data is clean, well-organized, and readily available for downstream processes.

Creating Online Tables

Online tables are the backbone for quick data retrieval in a RAG model. By creating these tables from the preprocessed data in the Delta table, we facilitate a seamless transition from raw data to actionable knowledge. Online tables provide a structured format that enhances the speed and efficiency of data retrieval, making them an essential component in the RAG pipeline.

Embedding Conversion

An essential function of embeddings is to convert formatted data to a form that generative models can readily handle. The data's semantic substance is captured throughout the conversion process by converting the online tables into embeddings. After that, they are saved in a specific repository so the RAG model may use them to generate responses.

Response Generation

The response generation phase leverages the embeddings stored in the repository to generate accurate and contextually relevant answers. The RAG framework can provide precise and informative responses by integrating retrieval mechanisms with generative models, significantly enhancing the user experience.

Methodology

The methodology adopted in this research includes the following steps:

1. **Data Loading and Preparation**: Data is loaded from a data warehouse and ingested into Databricks. The data is then cleaned and structured into a Delta table.

   The above image 1. illustrates the process beginning by loading data from a data warehouse into the Databricks environment, followed by ingestion and loading into a Delta table, ensuring the data is structured and clean.

2. **Online Table Creation**: The preprocessed data in the Delta table is converted into an online table to facilitate efficient data retrieval.

   The image continues to show the creation of an online table from the Delta table to facilitate quick and efficient data retrieval.

3. **Embedding Conversion**: The online table is transformed into embeddings, capturing the semantic meaning of the data.

   As depicted in the image, the online table is converted into embeddings that capture the semantic essence of the data.

4. **Embedding Storage**: The embeddings are stored in a repository and ready for use in response generation.

   The image illustrates that the embeddings are stored in a dedicated repository for future utilization by the RAG model.

5. **Response Generation**: The RAG model uses the stored embeddings to provide precise and situationally appropriate replies.

   The graphic concludes by showing how the RAG model uses embeddings to provide correct and relevant replies to the situation.

Options and Best Strategies Using Databricks

To effectively utilize Databricks in handling data warehouses as knowledge bases in RAG models, several strategies and options should be considered:

1. **Delta Lake for Data Management**: To effectively handle massive amounts of data, use Delta Lake. With Delta Lake, you can do ACID transactions, manage metadata at scale, and process streaming and batching data in one place.
2. **Optimized Data Ingestion**: Leverage Databricks' optimized data ingestion techniques to handle real-time and batch data loads seamlessly. This ensures that the data is always up-to-date and ready for use in RAG models.
3. **Automated ETL Pipelines**: Implement automated ETL (Extract, Transform, Load) pipelines to streamline data preparation. Databricks allow for the creation of robust ETL workflows that can be scheduled and monitored easily.
4. **Scalable Compute Resources**: Use Databricks' scalable compute resources to handle large-scale data processing tasks. This ensures that the data preparation and embedding conversion processes are performed efficiently.
5. **Collaborative Development Environment**: Use Databricks' collaborative development environment to allow data engineers, data scientists, and analysts to work together seamlessly. This fosters innovation and ensures that the best practices are followed.
6. **Advanced Analytics and Machine Learning**: Integrate advanced analytics and machine learning capabilities within Databricks to enhance the embedding conversion process. This can lead to more accurate and meaningful embeddings, improving the performance of the RAG model.
7. **Security and Compliance**: Use Databricks' integrated security capabilities to fulfil all data security and compliance needs. Compliance with industry standards, encryption of data, and access restrictions are all part of this.

Results

Integrating data warehouses as a knowledge base in a RAG model has led to notable improvements in response accuracy and retrieval speed. By employing a structured approach to data preparation in Databricks, including the creation of online tables and conversion into embeddings, the system has shown enhanced performance in generating precise and contextually relevant responses. The use of Delta tables for data management and optimized data ingestion techniques contributed to the efficiency of the process. Overall, the structured methodologies and embedding techniques have proven effective, showcasing the potential of RAG models in leveraging large data repositories for improved AI-driven decision-making.

Discussion

According to the results of this study, an adequate knowledge base may be created by merging data warehouses with RAG models. Structured data must undergo embedding conversion and preparation procedures to guarantee effective retrieval and precise answer creation. Customer service, data analysis, and BI are just a few areas that might benefit from this method.

Conclusion

Improve information retrieval and response generation using a Retrieval-Augmented Generation (RAG) model that employs data warehouses as a knowledge base. This study highlights the importance of structured approaches and embedding techniques in using massive datasets. We have shown significant improvements in response precision and retrieval velocity by preprocessing data in Databricks, generating online tables, and transforming these tables into embeddings.

Data is cleaned, organized, and easily retrievable when automated ETL pipelines, streamlined data intake, and Delta tables for data management are all in place. Improving the RAG model's accuracy and relevance to context is made possible by converting online tables into embeddings, which contain the data's semantic substance.

Databricks' collaborative development environment and scalable computing capabilities make data preparation and embedding conversion more efficient. In addition, the RAG model's overall performance is enhanced by the platform's powerful analytics and machine learning capabilities, which help create high-quality embeddings.

Benefiting greatly from AI-driven decision-making, this method has many potential applications, such as customer assistance, data analysis, and corporate intelligence. With Databricks, you can be confident that data is secure and that comply with all industry regulations.

The use of more sophisticated generative models to further enhance performance may be investigated in future studies, as can the application of this technique across several fields. By improving and expanding upon these methods, RAG models will become an essential resource for tapping into massive data stores in various contexts.

References

1. Databricks Documentation. (2023). Data Preparation in Databricks.
2. Lewis, Patrick & Perez, Ethan & Piktus, Aleksandara & Petroni, Fabio & Karpukhin, Vladimir & Goyal, Naman & Küttler, Heinrich & Lewis, Mike & Yih, Wen-tau & Rocktäschel, Tim & Riedel, Sebastian & Kiela, Douwe. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
3. J. Johnson, M. Douze and H. Jégou, "Billion-Scale Similarity Search with GPUs," in IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535-547, 1 July 2021, doi: 10.1109/TBDATA.2019.2921572
4. Dibouliya, Ashish. (2023). Review on: Modern Data Warehouse & how is it accelerating digital transformation.
5. Al-Okaily, A., Al-Okaily, M., Teoh, A. P., & Al-Debei, M. M. (2022). An empirical study on data warehouse systems effectiveness: the case of Jordanian banks in the business intelligence era. EuroMed Journal of Business. https://doi.org/10.1108/emjb-01-2022-0011

6. March, Salvatore & Hevner, Alan. (2007). Integrated decision support systems: A data warehousing perspective. Decision Support Systems. 43. 1031-1043. 10.1016/j.dss.2005.05.029.

Website:

1. https://encord.com/blog/retrieval-augmented-generation-rag-definition/
2. https://www.larksuite.com/en_us/topics/ai-glossary/data-warehouse