Elevating AI Applications with OpenSearch's Flow Framework and RAG Tool

Mingshi Liu
Owais Kazi

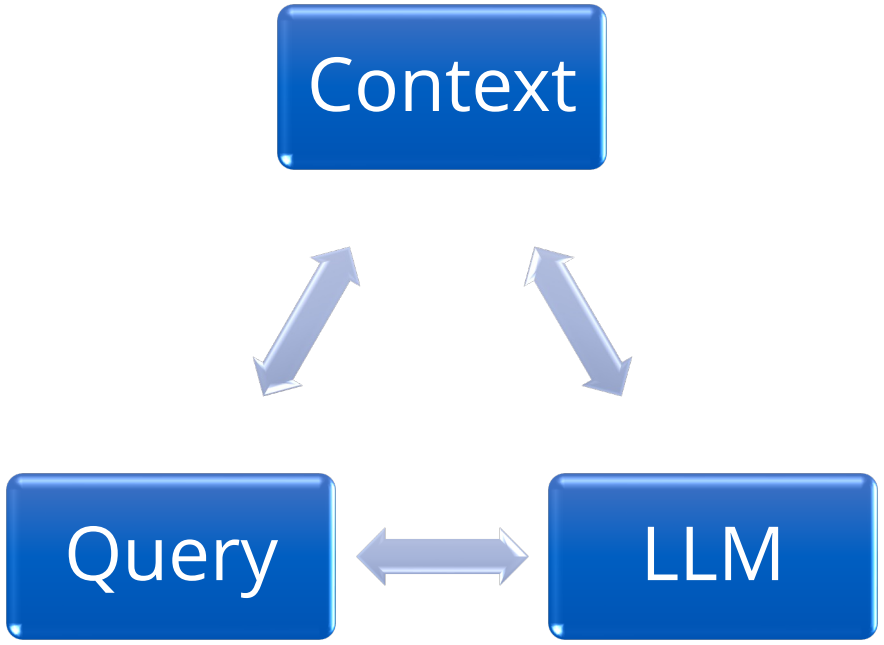
Agenda

- Introduction to RAG Tool
- Current ML Setup for Conversational Search
- Introduction to Flow Framework
- Demo
- Additional Default and Sample Use Cases

OpenSearch

What is RAG?

Retrieval-Augmented Generation



OpenSearch

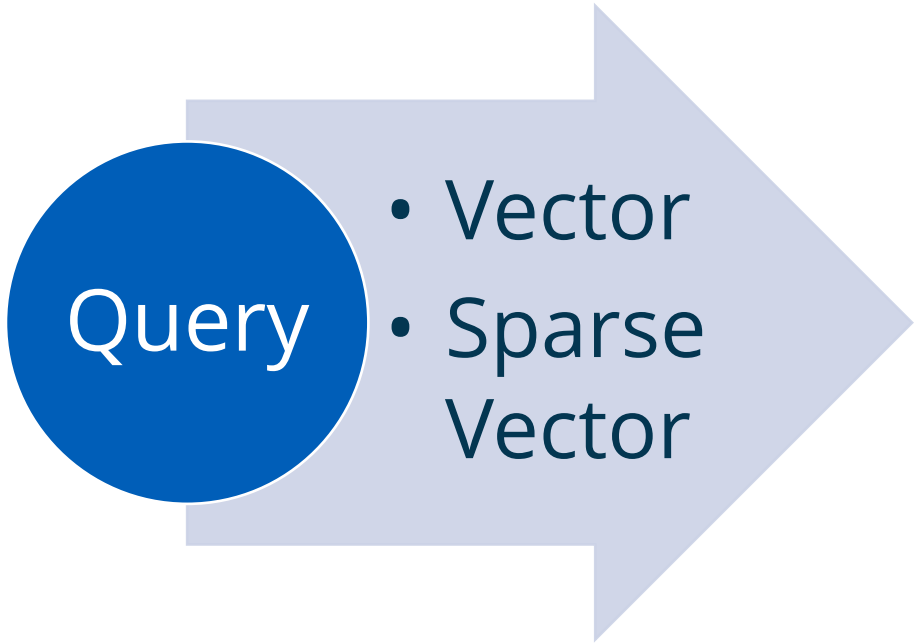
Introduction to RAG Tool

Retrieval-Augmented Generation

- Improve the quality of LLM generated response based on limited trained data
- Ensure the model can receive most recent and accurate data from search context (information retrieval)
- Reduce the needs of frequently training models

Introduction to RAG Tool

Retrieval-Augmented Generation

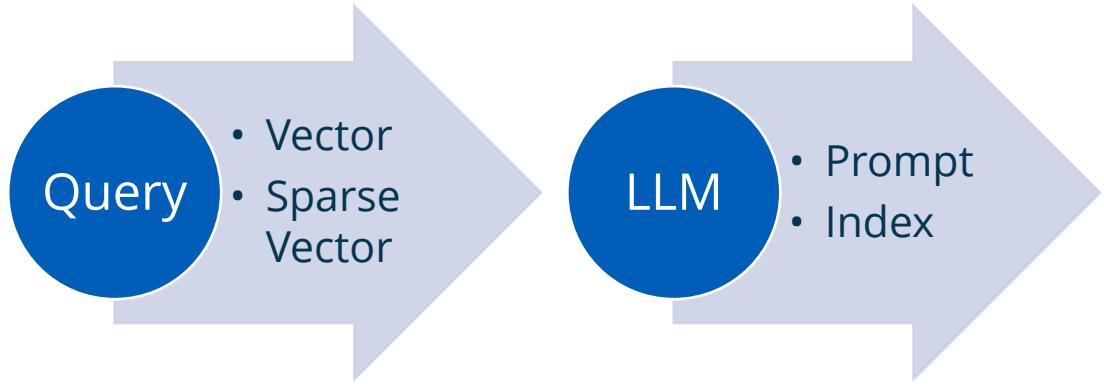


- Use vector query or sparse encoded query to index relevant data
  - Neural search
  - Neural sparse search

OpenSearch

Introduction to RAG Tool

Retrieval-Augmented Generation



- Use vector query or sparse encoded query to index relevant data
- Apply trained LLM (large language models) with prompt to summarize answers based on pretrained data and indexed data

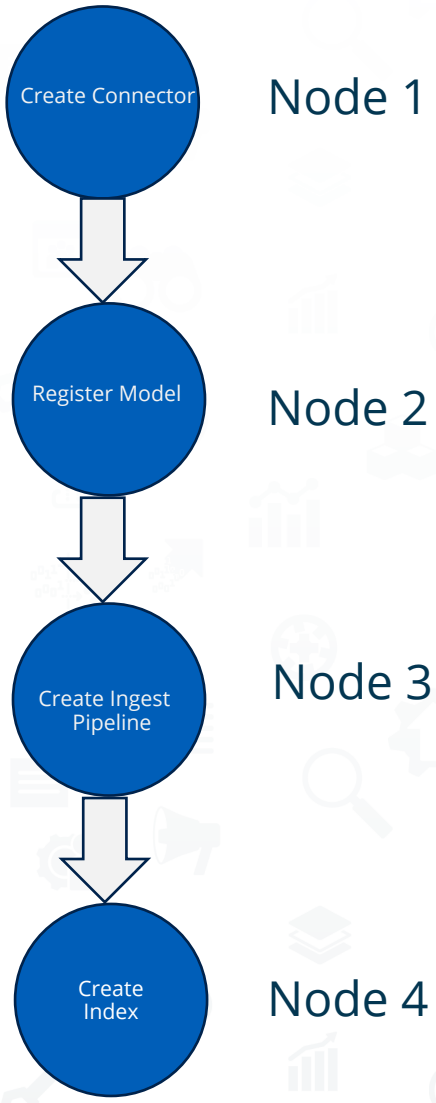
OpenSearch

Example: Current ML Setup for Conservational Search

- Create a connector for a remote model with pre and post functions
- Register an embedding model using the connectorID obtained from the previous step
- Configure an ingest pipeline to generate vector embeddings using the modelID of the registered model
- Create a k-NN index and add the pipeline created above
- Create a flow agent with RAGTool

Introduction to Flow Framework

Flow Framework



Node 1

Node 2

Node 3

Node 4

OpenSearch

Flow Framework

- Framework designed to streamline the cumbersome process of setting up complex ML use cases of OpenSearch
- Simplifies the complex setup with one click
- Provides automated templates
- Eliminates users to navigate the complexities of calling multiple APIs and waiting for their responses

OpenSearch

Demo

Additional Default and Sample Use Cases

- Explore more default use cases at here, with their corresponding defaults stored here

- Tailor templates according to your requirements. Sample templates are available here, and refer to our documentation here for further guidance.

OpenSearch

Thanks!
Any questions?

Owais Kazi

kazabdu@amazon.com

Mingshi Liu

minshil@amazon.com