Retail Insights Assistant

GenAI + RAG + Azure OpenAI + Multi-Agent System

AI-powered retail analytics for conversational Q&A & automated summarization

Problem Statement

Retail businesses generate **large volumes of data** (CSV, Excel, JSON, TXT).
Decision-makers want **instant insights** using **natural language queries**, like:

“Which category grew the most in Q3 in North region?”

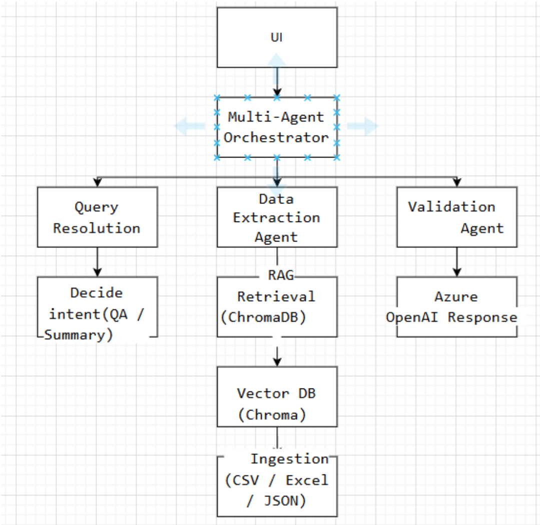
Challenges

- Manual analysis is slow
- Requires data engineering + AI
- Scaling beyond **100GB+** is hard

Goal

- Conversational Analytics (Q&A)
- Automatic Summarization
- Azure OpenAI Integration

Solution Architecture



Flow

User Query → Query Agent → RAG → Validation Agent → Azure OpenAI → Final Answer

Multi-Agent Architecture

Agent	Role
Query Resolution Agent	Understand user query, decide *mode* (Q&A / Summary), extract keywords
Data Extraction Agent (RAG)	Retrieve relevant chunks from vector DB
Validation Agent	Generate final answer using Azure OpenAI

LLM Integration (Azure OpenAI)

Why Azure?

Enterprise-grade
High context window (128K)
Pay-as-you-use

Models Used

Model	Purpose

`gpt-4o-mini`	Query resolution & answer generation
`text-embedding-large`	Embeddings for vector store (Chroma)

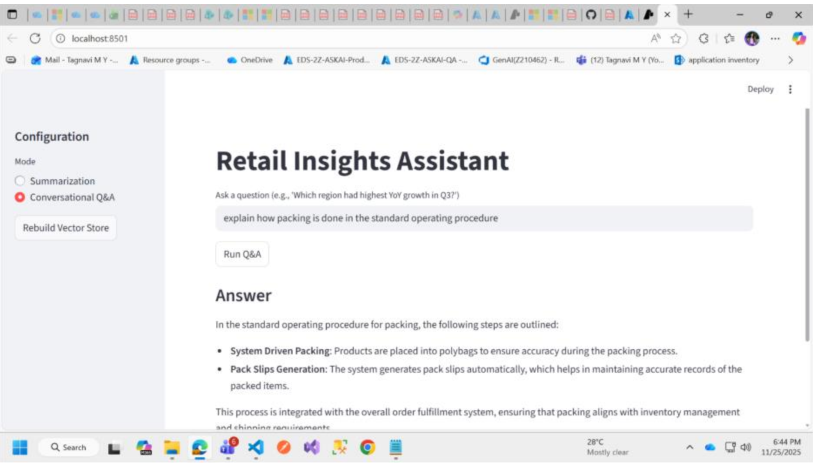
Data Storage & Indexing (100GB+ Design)

Component	Scalable Option
File Storage	Azure Data Lake / Blob Storage
Processing	PySpark / Azure Data Factory / Databricks
Indexing	ChromaDB / FAISS / Azure Cognitive Search
RAG	Azure Embeddings + Chroma

Key Strategy

- Chunk CSV / Excel files into text segments
- Convert → embeddings using Azure OpenAI
- Store in **persistent Chroma vector DB**
- Retrieval is `Top-K Semantic Search`

Query → Response Pipeline



Configuration

Mode

○ Summarization

● Conversational Q&A

Rebuild Vector Store

Retail Insights Assistant

Ask a question (e.g., 'Which region had highest YoY growth in Q3?')

explain how packing is done in the standard operating procedure

Run Q&A

Answer

In the standard operating procedure for packing, the following steps are outlined:

- **System Driven Packing**: Products are placed into polybags to ensure accuracy during the packing process.
- **Pack Slips Generation:** The system generates pack slips automatically, which helps in maintaining accurate records of the packed items.

This process is integrated with the overall order fulfillment system, ensuring that packing aligns with inventory management and shipping requirements

Pipeline:

```
1) Query Resolution Agent → detects mode = Q&A
2) Normalizes query
3) RAG retrieves relevant chunks from vector DB
4) Validation Agent sends docs + query to Azure GPT
5) GPT produces clean business answer
```

Cost & Performance Optimization

Feature	Benefit
Vector store persistence	Avoid re-ingestion
Top-K retrieval	Limits token usage
Prompt compression	Smaller context

Final Summary

The Retail Insights Assistant is a GenAI-driven system that enables conversational analytics on retail data using Azure OpenAI, RAG, and a multi-agent architecture. It ingests structured and unstructured files, stores them in a vector database, and generates summaries or answers business queries using natural language.