Knowledge Discovery Engine: A Full-Stack Scientific Search System

Unilever

Outline

- Why did we do this?
- Requirements
- System components & design
- GenAI
- Q&A

About

John Labarga

Senior Manager, Data Science and Machine Learning

Unilever, Digital & Partnerships

Joined Unilever in 2021

Previously: Sanofi, Lockheed Martin

Founded 1929

Top 5 FMCG/CPG by revenue

Owns brands like Dove, Axe, Knorr, Ben & Jerry's, Hellman's

Strategically invested in sustainability and digital

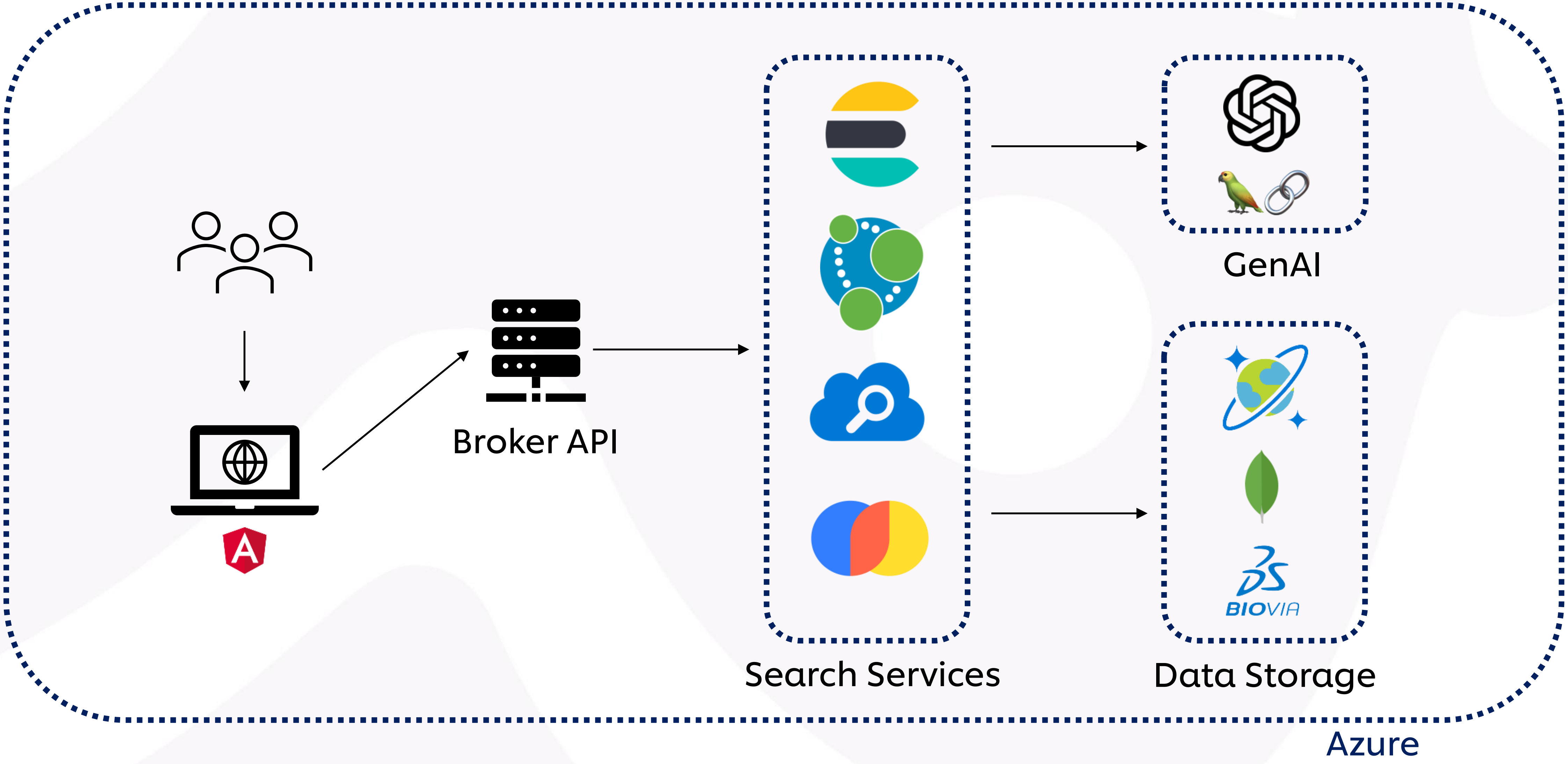
Why?

- Scientific research is a core activity in CPG industry
- Unilever produces a large volume of lab reports, experimental notebooks, etc.
- Highly-distributed organization
- Insights from one product type often applicable to another

Why?

- Commercial off-the-shelf (COTS) systems mostly don't offer fusion of public and private sources
- Search and NLP are fast-moving areas, need to quickly build and deploy extensions/new search modes
  - E.g. RAG went from obscure to table stakes in < six months
- Scientists need to be able to fine-tune and customize performance
- Information is highly sensitive, all components must be hosted in our Azure environment

System Design



Components: UI

- Different users prefer different interaction modes:
  - Keyword search
  - Semantic search
  - RAG and chat-with-your-document(s) (BYOD)
  - Agent chatbot

- Term completion useful in chemistry
  - Query custom embedding models and knowledge graph entities

How can we help you?

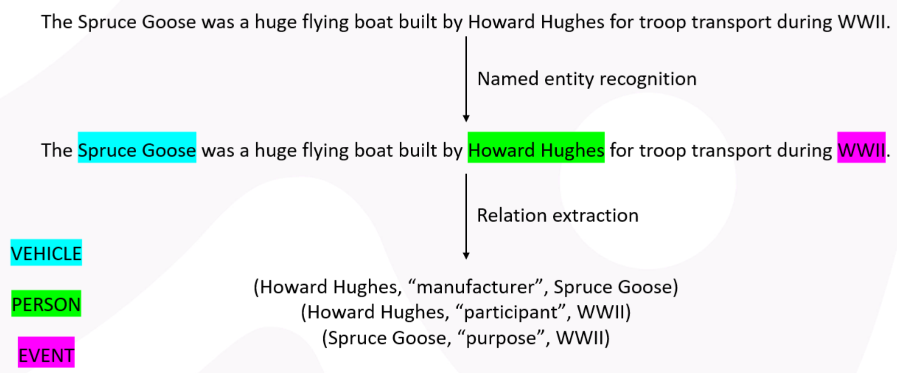
Components: Knowledge Graph

- **Things present in the knowledge graph need to be facts from authoritative sources**

- **General procedure for constructing a knowledge graph:**

  1. **Segment document into spans such that a fact won't generally transcend a span (paragraphs usually work)**
  2. **Run entity recognition and coreference resolution**
  3. **Normalize entities**
  4. **Run relation extraction on entity pairs**

Extracting knowledge from a short sentence:



The Spruce Goose was a huge flying boat built by Howard Hughes for troop transport during WWII.

Named entity recognition

The Spruce Goose was a huge flying boat built by Howard Hughes for troop transport during WWII.

Relation extraction

VEHICLE

PERSON

EVENT

(Howard Hughes, "manufacturer", Spruce Goose)
(Howard Hughes, "participant", WWII)
(Spruce Goose, "purpose", WWII)

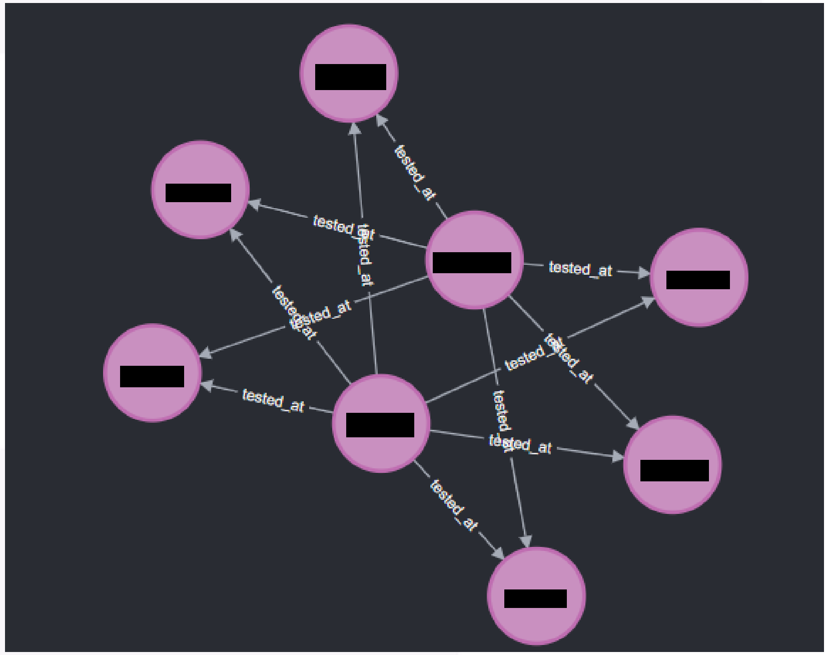
Components: Knowledge Graph

- **spaCy very capable at named entity recognition**
  - **Models available for many domains (e.g. scispaCy)**
  - **Fine-tunable for custom entity types, and supports GPU acceleration**

- **Chose [OpenNRE] for relation extraction**
  - **Supports fine-tuning and GPU acceleration**
  - **Reliable confidence scores, useful for filtration**

- **GPT-4 can do relation extraction fairly well**
  - **GPT-3.5 not much better than OpenNRE, cost not justifiable**

Finding products tested at the same sites:



Components: Knowledge Graph

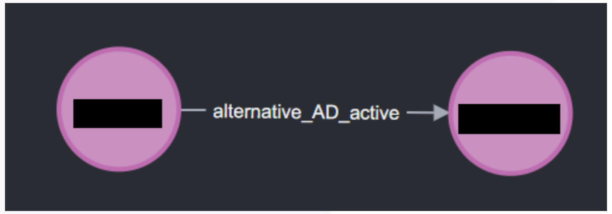
- Constructing entire knowledge graph with GPT-4 is very expensive

- Bootstrap from GPT-4 instead:
  1. Identify a high-quality, diverse sub-corpus
  2. Write extraction prompts, potentially specific to document types
  3. Include few-shot examples
  4. Run extraction on “golden set” from step 1
  5. Tune locally-hostable NER and RE with examples
  6. Run rest of extraction job with these tuned models

Components: Search

- Classic Boolean keyword search powered by ElasticSearch or Azure Cognitive Search (ACS)
- Semantic search: embed document chunks and user query, use FAISS for similarity
  - OpenAI and HuggingFace embeddings both work well
  - Many scientists want to search in a terminology-agnostic way, not RAG per se
- Directly search the knowledge graph
  - E.g. what chemical of type X is used in the same reactions as chemical Y of type Z?

Finding a specific relationship between two chemicals:

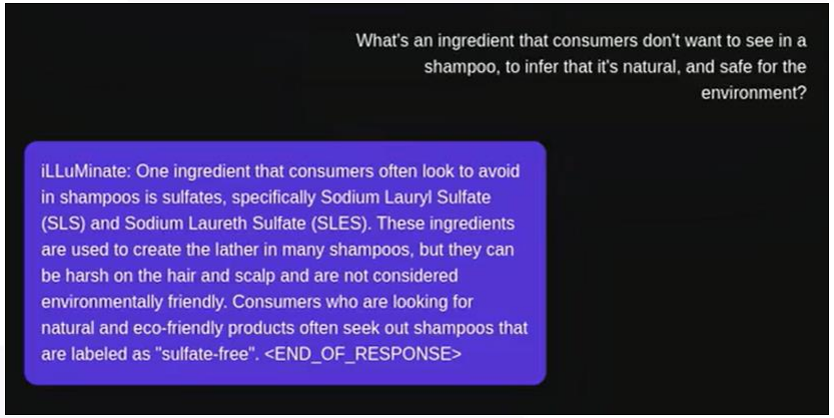


Retrieval-Augmented Generation (RAG)

- Document vectors stored in compartmentalized Azure Cognitive Search instances
- Interaction workflow designed in Prompt Flow
- During RAG, can select a particular document to “chat with”
    - Full text is loaded into prompt, user drops down into a separate window
    - Enables isolate -> interrogate workflow for deep dives

Agent

- **Controller + tool agent written in LangChain**
- **Direct connection to ACS as vector store for internal documents, ChromaDB for public documents, and Neo4j for knowledge graph querying**
- **Highly extensible, just add a tool to enable new functionality**
- **Prompting strategy easy to customize (tree-of-thoughts, prompt chain, etc.)**



What's an ingredient that consumers don't want to see in a shampoo, to infer that it's natural, and safe for the environment?

iLLuMinate: One ingredient that consumers often look to avoid in shampoos is sulfates, specifically Sodium Lauryl Sulfate (SLS) and Sodium Laureth Sulfate (SLES). These ingredients are used to create the lather in many shampoos, but they can be harsh on the hair and scalp and are not considered environmentally friendly. Consumers who are looking for natural and eco-friendly products often seek out shampoos that are labeled as "sulfate-free". <END_OF_RESPONSE>

Lessons learned

- It’s easy to build this whole thing yourself!
- LLMs capable of supplanting many older tools, but costs need to come down
- For technical agents, try to model the expert’s thought process
- Know how much you trust your sources