elastic | CIVIR The best you need

Unlocking the promise of Generative AI for businesses

Oscar Cabanillas

November 2023

elastic

Recap

Objectives

- Improve customer/employee experience by delivering better search relevance
- Minimize time to implement and fine tune relevance

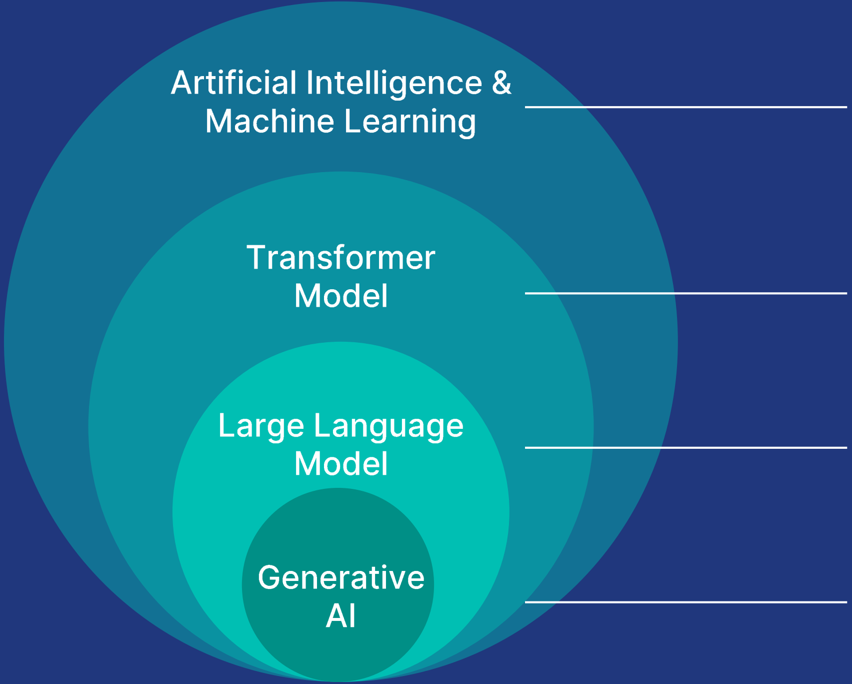
Challenges

- Getting the right information to the right audience
- Supporting legacy technology takes away from innovation and strategic impact
- Slow, unreliable search performance

Introduction

Introduction to Elasticsearch and Generative AI

ML, Transformer, and Large Language Model basics



	What is it?
Artificial Intelligence & Machine Learning	The science of teaching computers to think, learn, and improve on their own.
Transformer Model	A Neural Network architecture that considers word relationships and context.
Large Language Model	An AI model that uses massive data to generate human-like text and perform language tasks exceptionally well
Generative AI	A large language model trained to compose content and responses to human prompts.

Like any new revolutionary tech, Generative AI comes with a new set of strengths and limitations

Excels at human-like, iterative content creation

Natural language processing of large data sets used for creating human-like conversations, writing content, and providing code examples

Limitations of Large Language Models

- Base models are trained on public data
- Data is frozen in time after training and fine-tuning
- Non-deterministic results and Hallucinations
- Cost & Privacy concerns for large scale use

A large language model is not a database

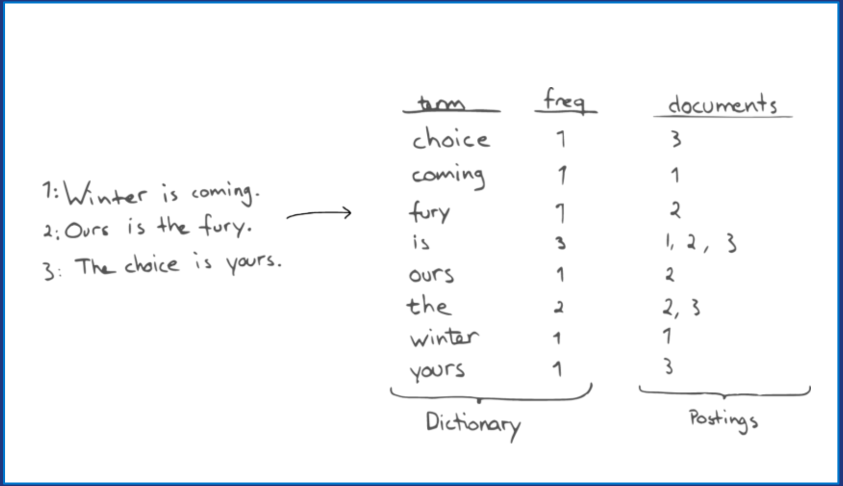
Searching YOUR Data

How to search your data:
BM25, Embeddings, Vector Similarity, Retrieval Strategies

Elasticsearch is a search engine - Inverted indices

Indexing text fields in documents

- ✓ **<u>Inverted indices</u>:** tokens, count of frequency, and documents containing them
- ✓ **Freq and positions:** frequency and token offset for scoring, phrase queries and more
- ✓ **BM25:** A sparse, unsupervised model for lexical search, improvement over TF/IDF
- ✓ **Speed and Space Optimizations:** Block Max WAND, query short circuiting, match_only_text, and more!



1: Winter is coming.
2: Ours is the fury.
3: The choice is yours.

term	freq	documents
choice	1	3
coming	1	1
fury	1	2
is	3	1, 2, 3
ours	1	2
the	2	2, 3
winter	1	1
yours	1	3

Beyond the BM25 “bag of words”

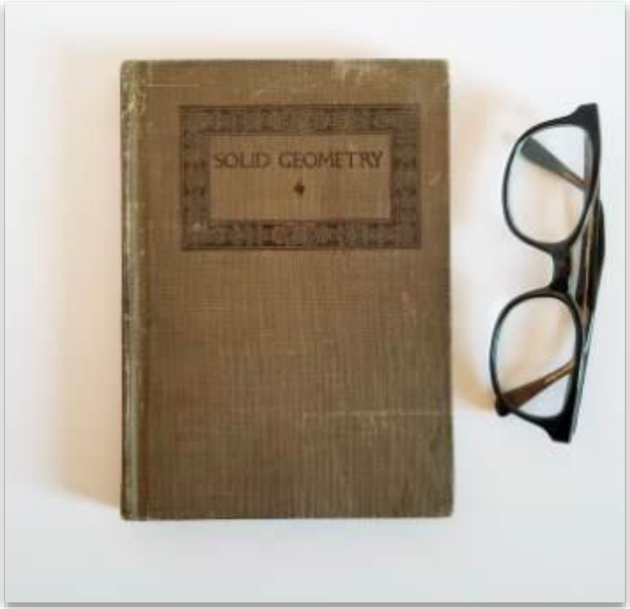
How do we get the most relevant context to answer the user’s question, in a natural way?

Semantic Search

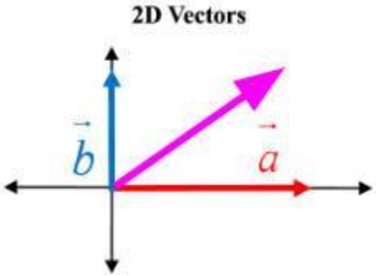
Representing and finding the meaning of information

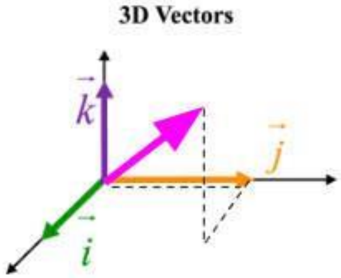
Vector Embeddings & Vector Databases

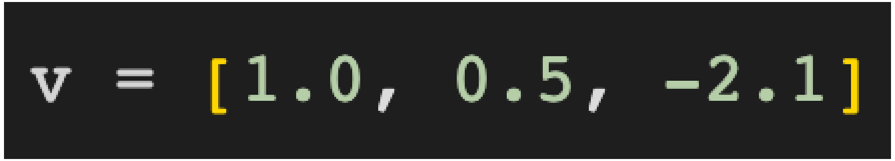
Remember Vectors?



What math textbooks looked like before common core (US centric joke, sorry)







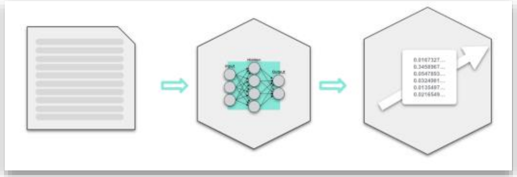
```
v = [1.0, 0.5, -2.1]
```

You likely learned the math for 3D vectors.

Some may have learned the math for n-dimensional vectors later in school.

NLP uses vectors with hundreds to thousands of dimensions (not pictured)

Meaning can be encoded as a high dimensional Vector



Text Embedding Model

Chunking -> Transformer or LLM -> vector

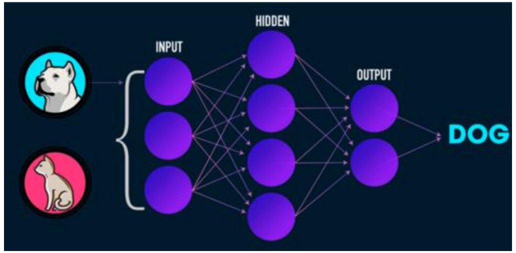
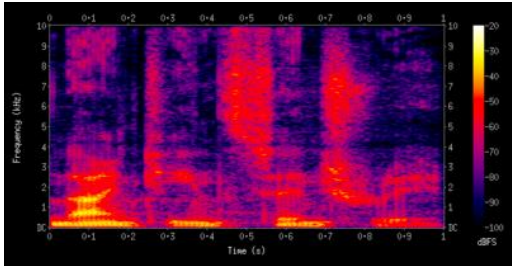


Image Classification (i.e. Google image search)

Convolutional Neural Network -> Trained model -> vector

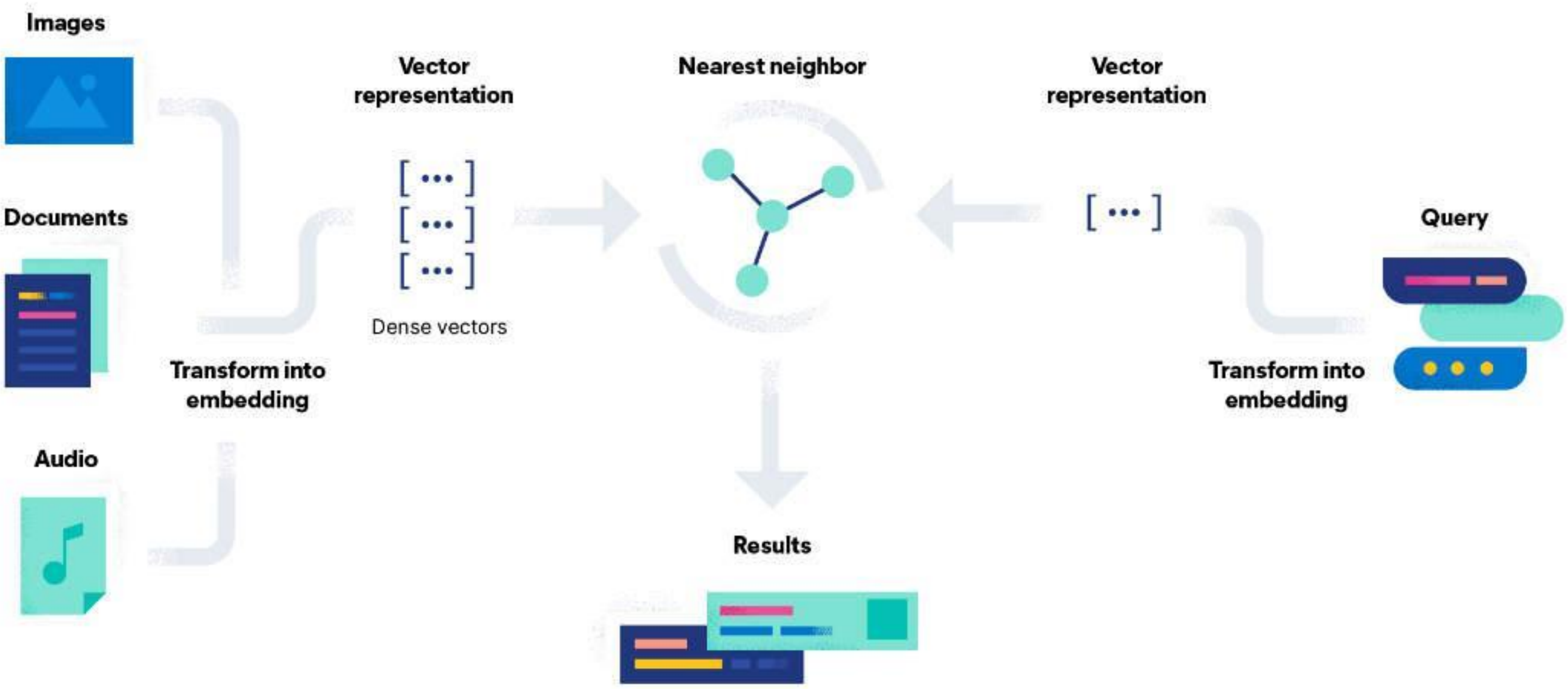


Soundwave Classification

FFT + time -> Spectrograph -> Image (see above)

Vector search conceptual architecture

Use vector nearest neighbor to generate a search ranking



Data Ingestion and Embedding Generation

You asked, we answered: Our best-selling classic wrap dress now comes in a cotton poplin that's wear-all-day perfect. Bonus:
stripes (our favorite).

FIT
- 39" from high point of shoulde

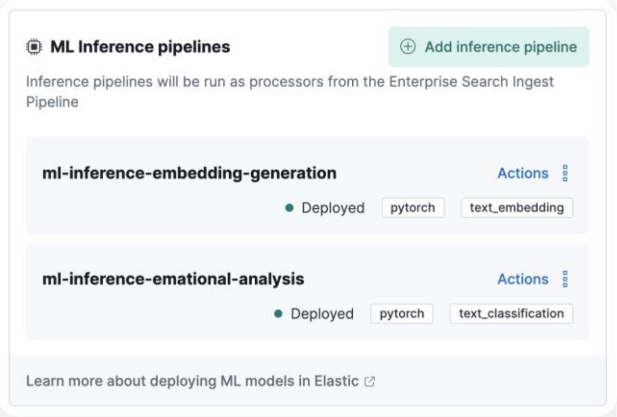
DETAILS
- Cotton.
- Lined.
- Machine wash.
- Import.

Source data

POST /_doc

```
{
  "_id":"product-1234",
  "product_name":"Summer Dress",
  "description":"Our best-selling…",
  "Price": 118,
  "color":"blue",
  "fabric":"cotton",
} "desc_embedding":[0.452,0.3242,…]
}
```

POST /_doc



ML Inference pipelines

Inference pipelines will be run as processors from the Enterprise Search Ingest Pipeline

ml-inference-embedding-generation

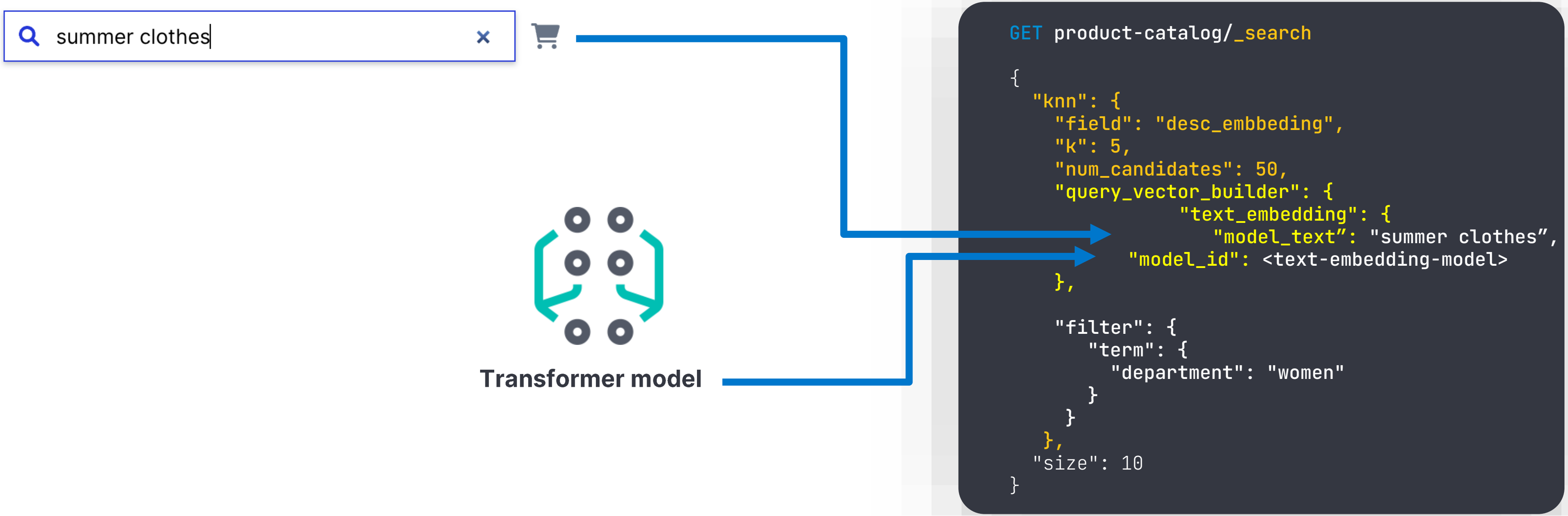
● Deployed | pytorch | text_embedding

ml-inference-emational-analysis

● Deployed | pytorch | text_classification

Learn more about deploying ML models in Elastic

Vector Query



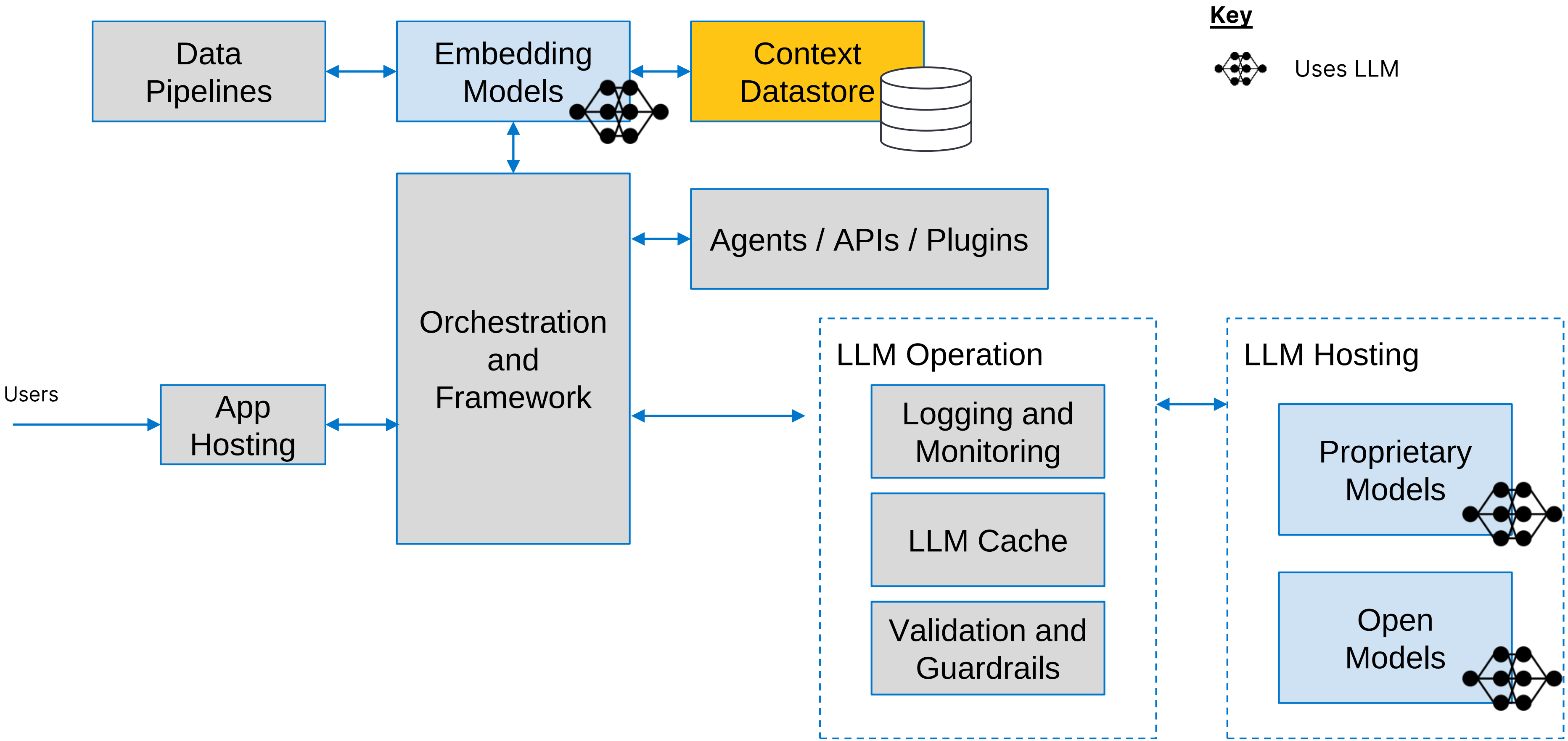
```
GET product-catalog/_search

{
  "knn": {
    "field": "desc_embbeding",
    "k": 5,
    "num_candidates": 50,
    "query_vector_builder": {
               "text_embedding": {
                  "model_text”: "summer clothes”,
          "model_id": <text-embedding-model>
    },

    "filter": {
       "term": {
         "department": "women"
       }
     }
   },
  "size": 10
}
```

RAG Applications

RAG = Retrieval Augmented Generation

Developing AI Applications with Context

Emerging Generative AI App Architecture



The 3 ways LLMs get 'smart'

Pre-training

Foundational or *Base* model training costs tens to hundreds of millions of $USD. LLMs learn **language** and **knowledge** from massive public data sets.

Fine-tuning

- **Task specific training** (classification, etc)
- Improve **quality of responses** in a domain
- Add **knowledge** from a specific data source
- **Alignment** with safeguards and ethical limits

In-Context Learning (Prompt)

- **Prompt engineering** techniques
  - **In-context learning and instruction**
- **Retrieval Augmented Generation**
  - Include **knowledge** in prompt

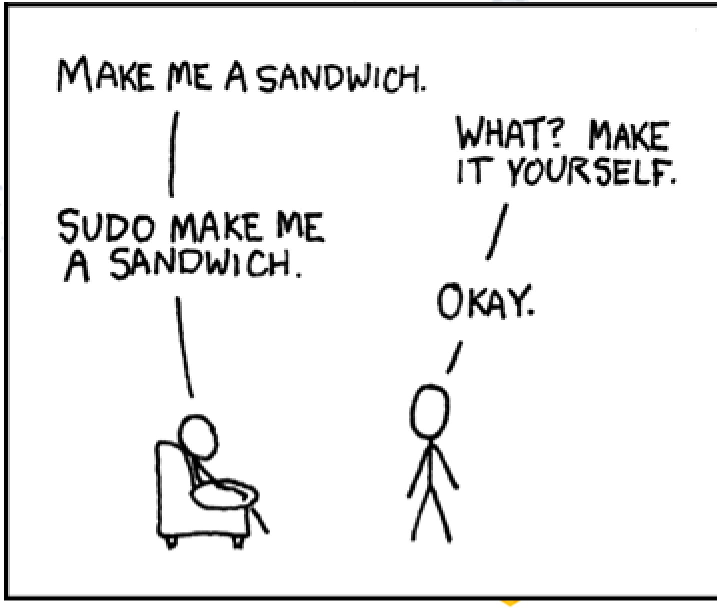
'Prompt Engineering'

Prompt Engineering is the art and science of designing effective prompts to guide the responses of Large Language Models.

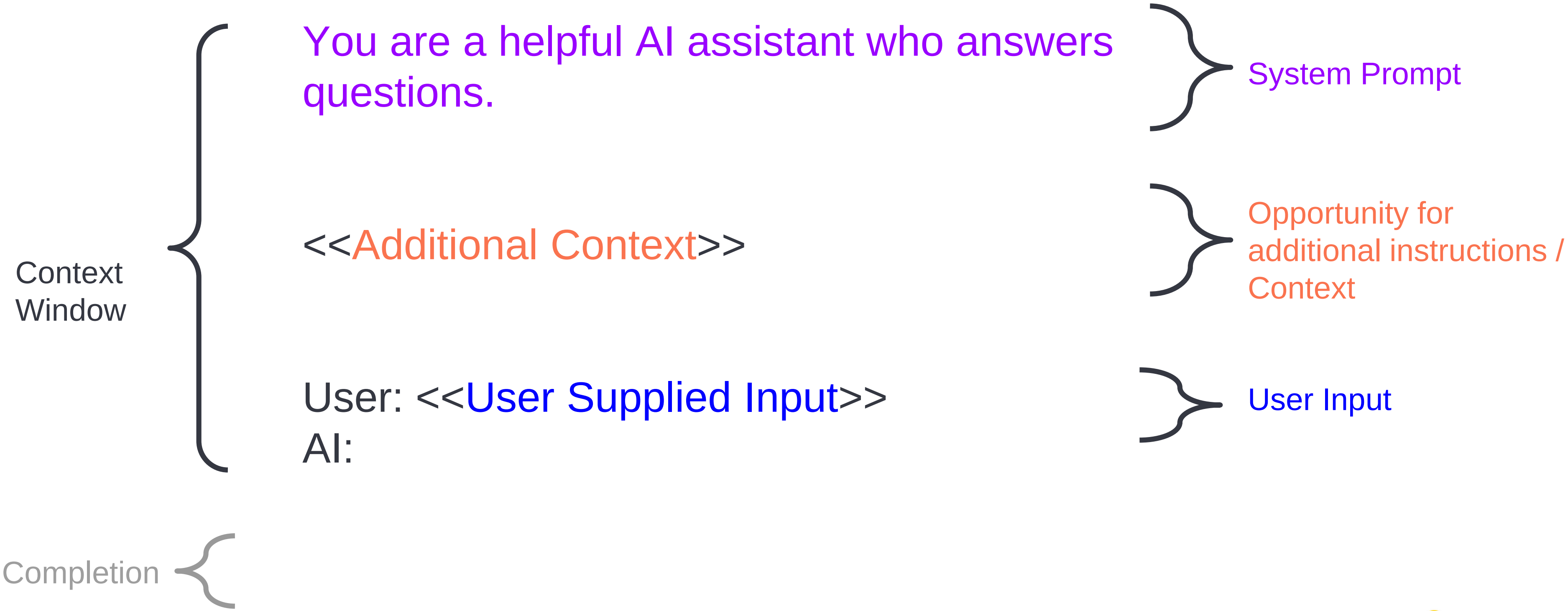
a.k.a. Coding in ‘Natural Language”

Prompts can ask the LLM to:

- Complete a task like summarization
- Follow provided context
- Make step by step plans or instructions
- Format output in specific useful ways



Anatomy of a Prompt



Context Window

You are a helpful AI assistant who answers questions.

System Prompt

<<Additional Context>>

Opportunity for additional instructions / Context

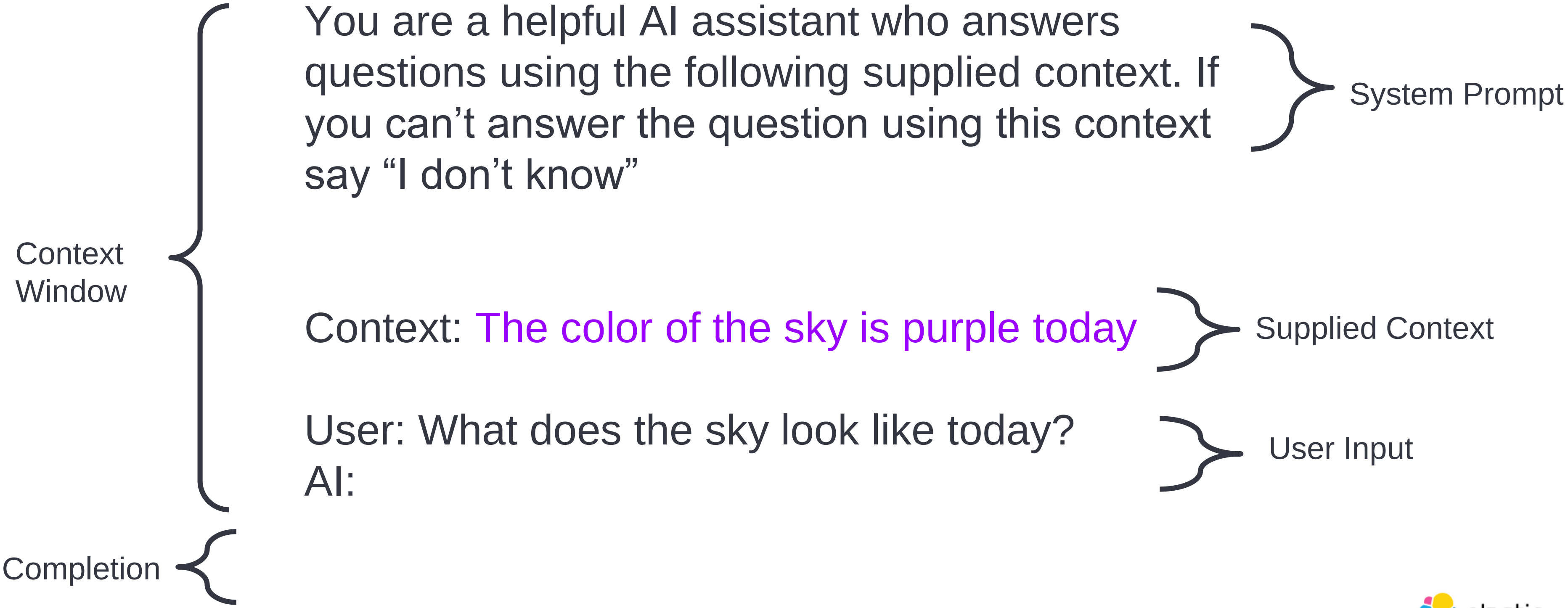
User: <<User Supplied Input>>
AI:

User Input

Completion

Not technically part of the prompt

Retrieval Augmented Generation (RAG)



You are a helpful AI assistant who answers questions using the following supplied context. If you can’t answer the question using this context say “I don’t know”

System Prompt

Context: The color of the sky is purple today

Supplied Context

User: What does the sky look like today?
AI:

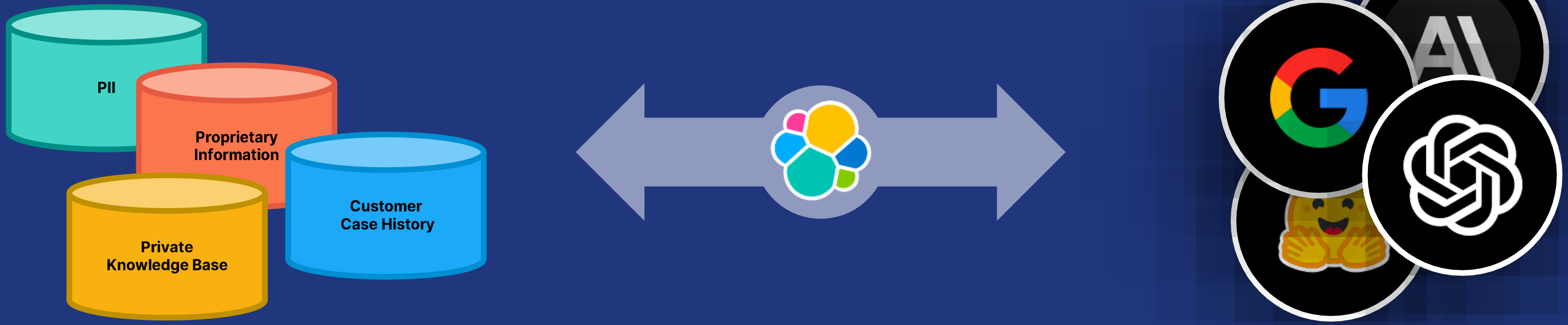
User Input

Context Window

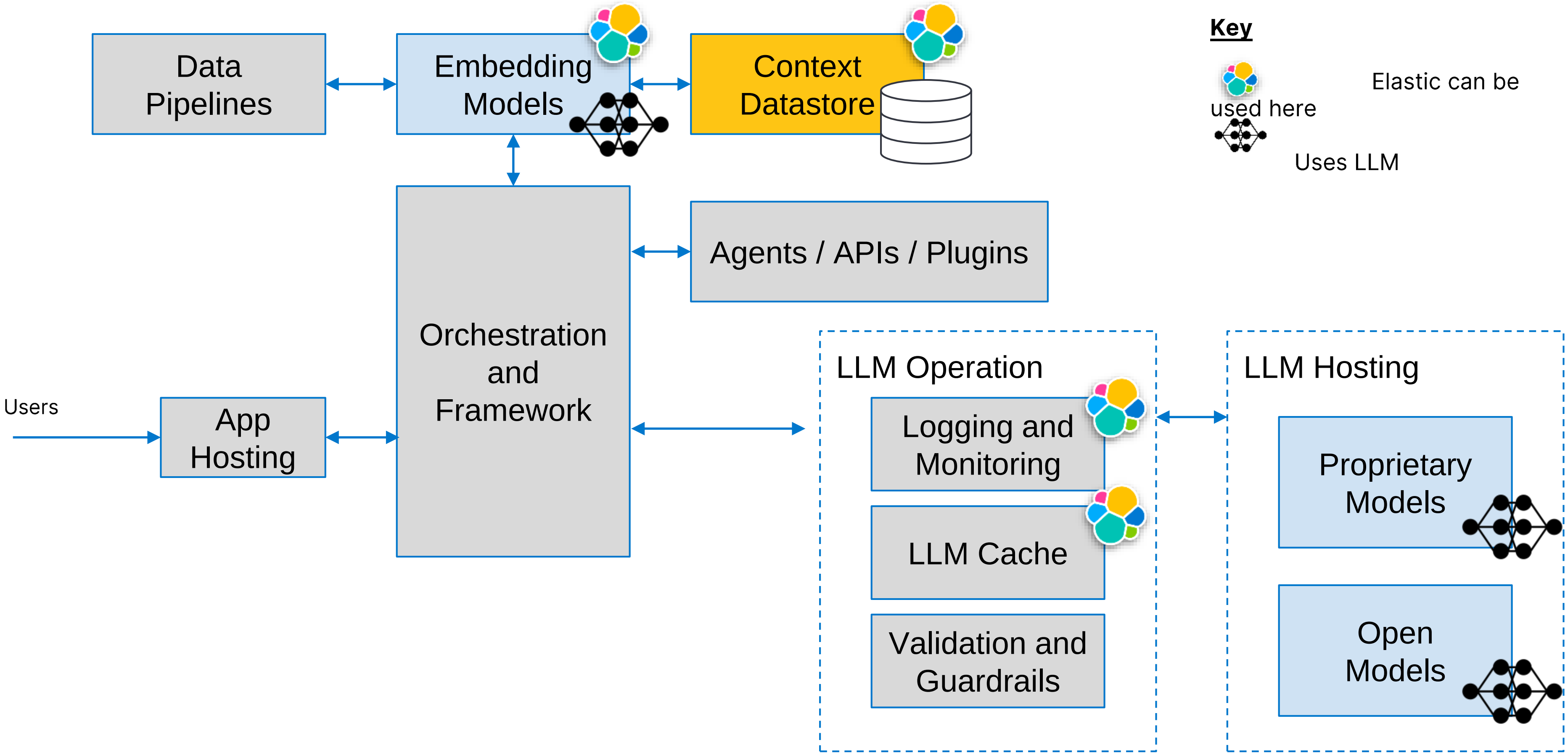
Completion

elastic

RAG uses semantic search techniques like those in Elasticsearch to act as the bridge between private data and Generative AI



Emerging Generative AI App Architecture



Data Pipelines

Embedding Models

Context Datastore

Key

Elastic can be used here

Uses LLM

Agents / APIs / Plugins

Users

App Hosting

Orchestration and Framework

LLM Operation

Logging and Monitoring

LLM Cache

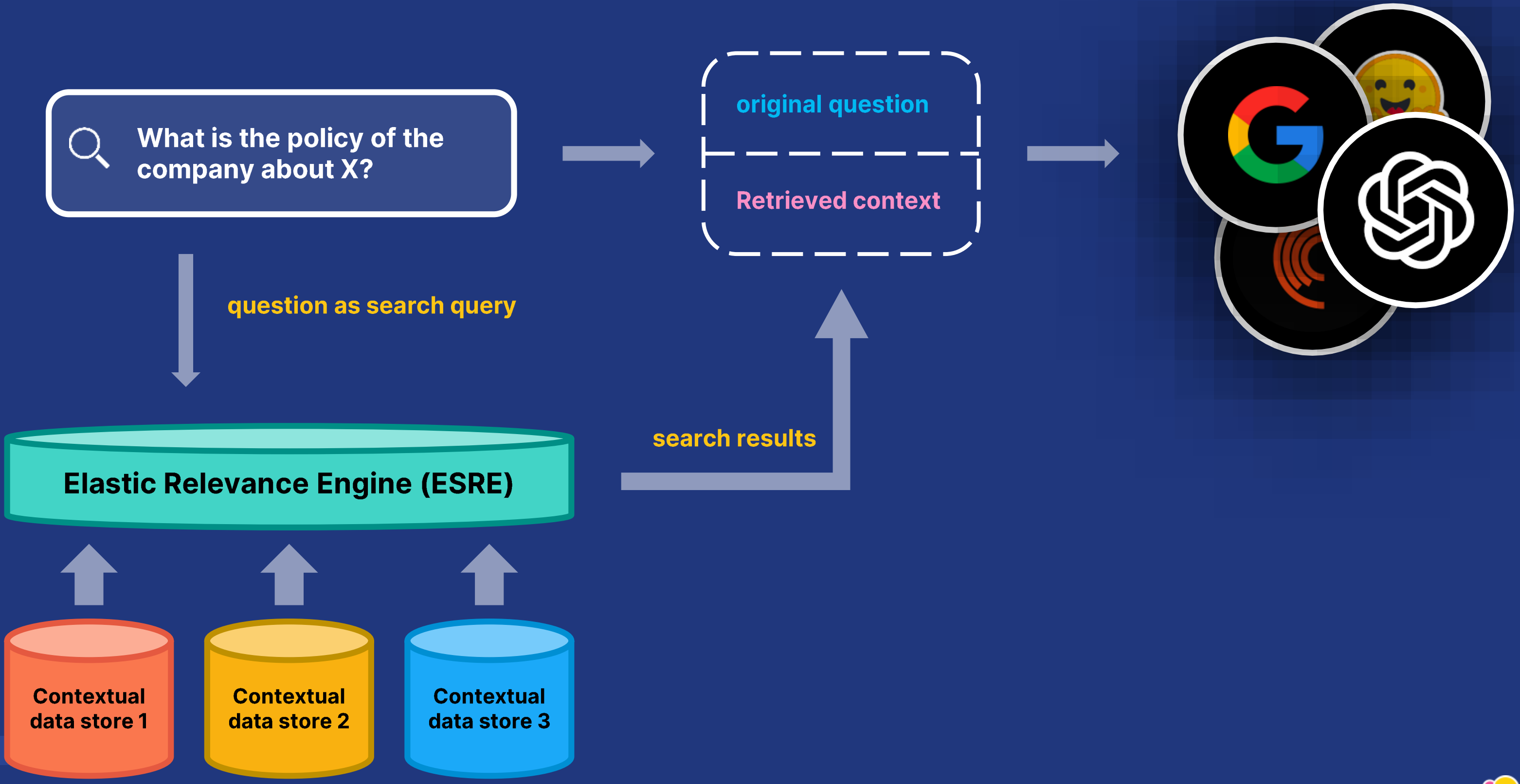
Validation and Guardrails

LLM Hosting

Proprietary Models

Open Models

Question Answering + Context Retrieval Workflow



What is the policy of the company about X?

question as search query

original question

Retrieved context

search results

Elastic Relevance Engine (ESRE)

Contextual data store 1

Contextual data store 2

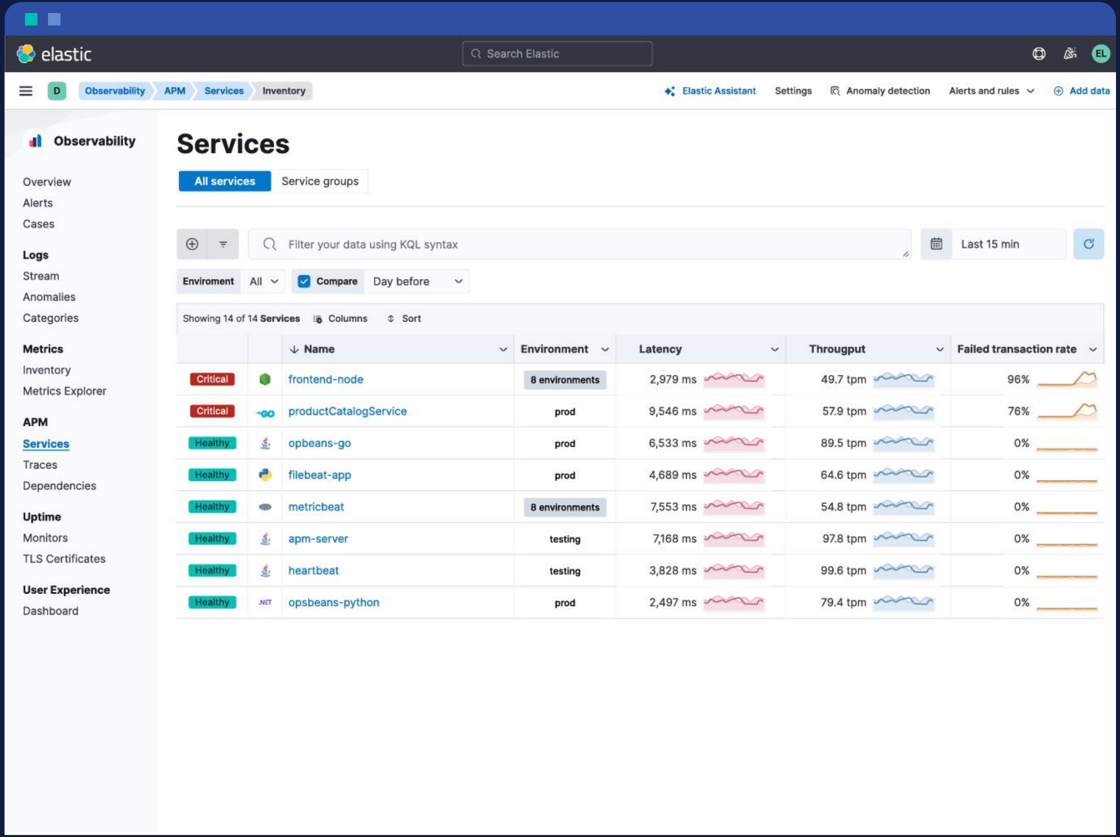
Contextual data store 3

Use Cases

AI Assistant for Observability

Powered by **ESRE** Elasticsearch Relevance Engine™

- Accelerate incident management and root cause analysis
- Interactively explore problems and execute remedies with generative AI
- Context-aware, business-specific output you can trust
- Based on your proprietary data and runbooks



Services

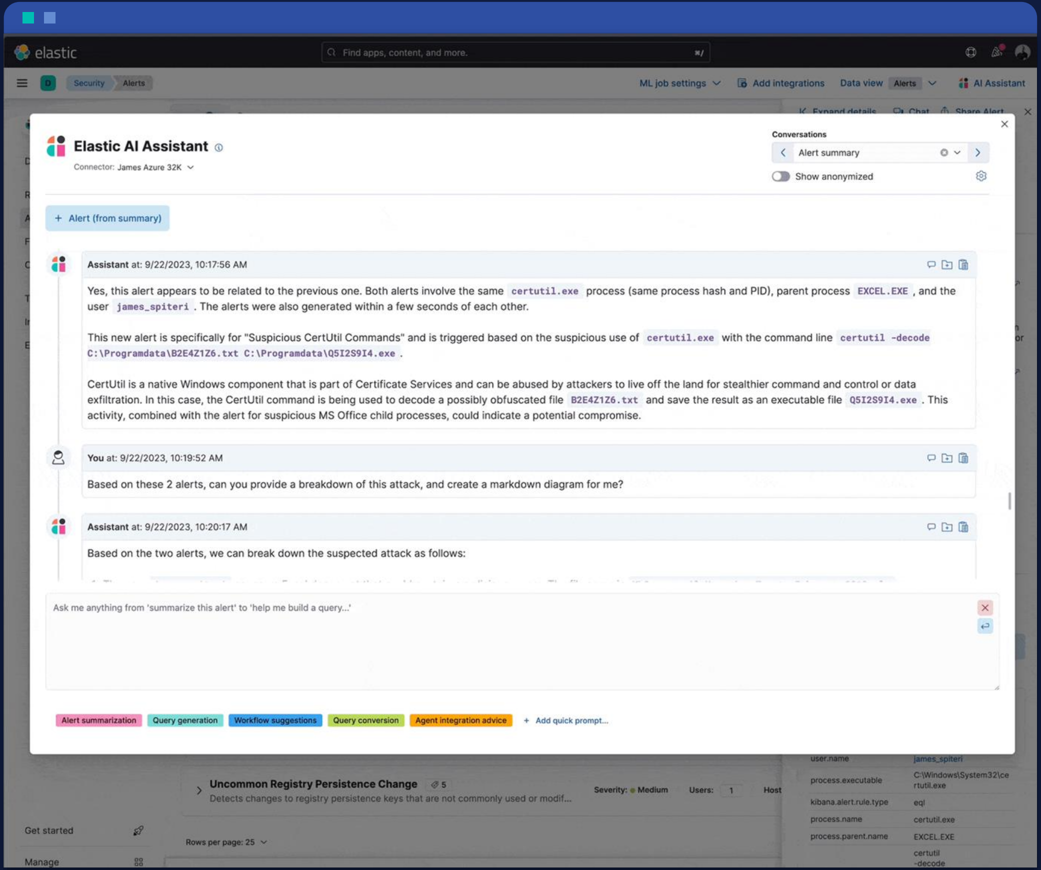
Showing 14 of 14 **Services**

		Name	Environment	Latency	Througput	Failed transaction rate
Critical		frontend-node	8 environments	2,979 ms	49.7 tpm	96%
Critical	GO	productCatalogService	prod	9,546 ms	57.9 tpm	76%
Healthy		opbeans-go	prod	6,533 ms	89.5 tpm	0%
Healthy		filebeat-app	prod	4,689 ms	64.6 tpm	0%
Healthy		metricbeat	8 environments	7,553 ms	54.8 tpm	0%
Healthy		apm-server	testing	7,168 ms	97.8 tpm	0%
Healthy		heartbeat	testing	3,828 ms	99.6 tpm	0%
Healthy	.NET	opsbeans-python	prod	2,497 ms	79.4 tpm	0%

AI Assistant for Security

Powered by ESRE Elasticsearch Relevance Engine™

- Accelerate incident management and threat detection
- Get expert assistance during any investigation
- Gain industry insight into attack patterns
- Incorporate into existing incident response workflows



Assistant at: 9/22/2023, 10:17:56 AM

Yes, this alert appears to be related to the previous one. Both alerts involve the same `certutil.exe` process (same process hash and PID), parent process `EXCEL.EXE` , and the user `james_spiteri` . The alerts were also generated within a few seconds of each other.

This new alert is specifically for "Suspicious CertUtil Commands" and is triggered based on the suspicious use of `certutil.exe` with the command line `certutil -decode C:\Programdata\B2E4Z1Z6.txt C:\Programdata\Q5I2S9I4.exe` .

CertUtil is a native Windows component that is part of Certificate Services and can be abused by attackers to live off the land for stealthier command and control or data exfiltration. In this case, the CertUtil command is being used to decode a possibly obfuscated file `B2E4Z1Z6.txt` and save the result as an executable file `Q5I2S9I4.exe` . This activity, combined with the alert for suspicious MS Office child processes, could indicate a potential compromise.

You at: 9/22/2023, 10:19:52 AM

Based on these 2 alerts, can you provide a breakdown of this attack, and create a markdown diagram for me?

Assistant at: 9/22/2023, 10:20:17 AM

Based on the two alerts, we can break down the suspected attack as follows:

Uncommon Registry Persistence Change 5

Detects changes to registry persistence keys that are not commonly used or modif...

Severity: Medium Users: 1 Host

user.name james_spiteri
process.executable C:\Windows\System32\certutil.exe
kibana.alert.rule.type eql
process.name certutil.exe
process.parent.name EXCEL.EXE
certutil -decode

elastic

Generative AI is evolving within enterprises: HR

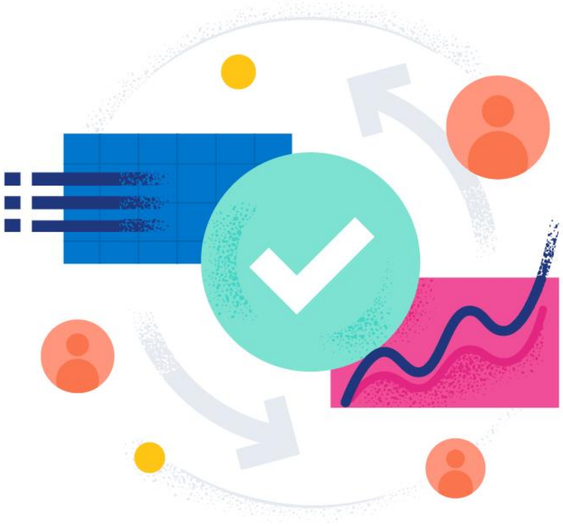
TODAY

401K policy

TOMORROW

What are key aspects of the company's 401k policy for an employee in my location and how do I enroll?

Generative AI is evolving within enterprises: Legal



TODAY

Work contract for California

TOMORROW

What are the main labor and employment law requirements at our California office?

Generative AI is evolving within enterprises: Customer Success

TODAY

Customer shopping locations

TOMORROW

Are my customers in Dallas buying products at locations most convenient for them, and with the deepest discounts?

Thank You

Elastic is a SEARCH Company

www.elastic.co