Red Hat Summit

Connect

# Sviluppo di soluzioni Agentic AI con Elastic

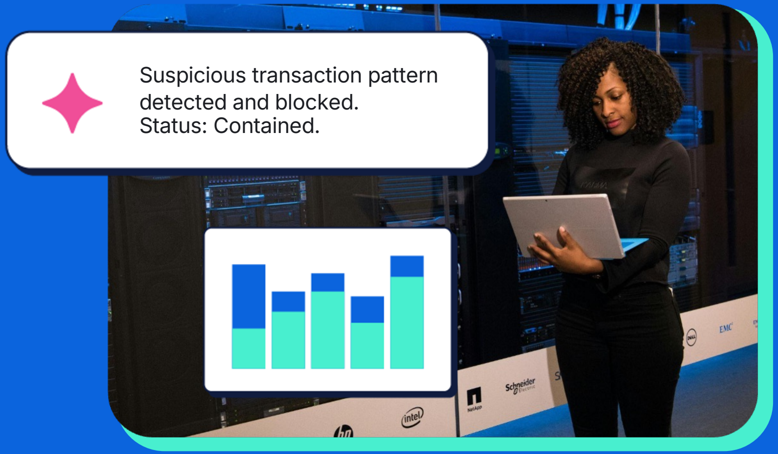
Maurizio Romani – Senior Solutions Architect

elastic | Red Hat

# AI is here. And it’s changing everything.

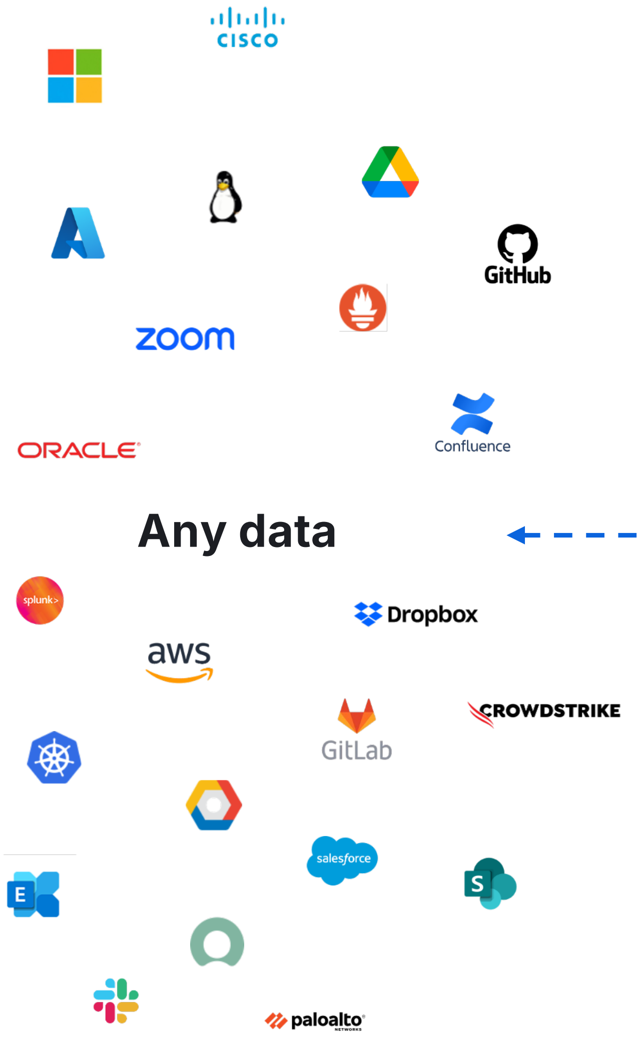
## In the workplace, in our personal lives, and more ...

**People** driving the process with AI as assistants & copilots



Suspicious transaction pattern
detected and blocked.
Status: Contained.

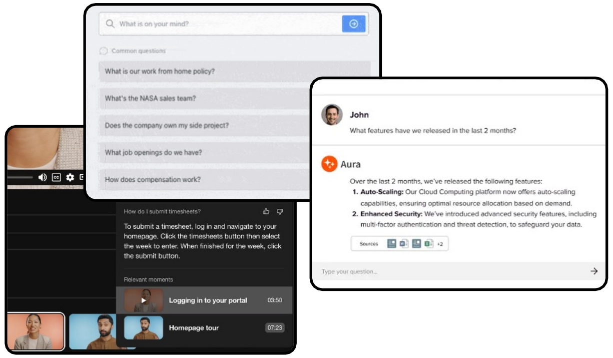
**AI agents** driving the process with people reviewing



## Any data

Success depends on providing AI **high-quality, relevant** access to **your data and powerful tools**

## AI experiences


What is on your mind?
Common questions
What is our work from home policy?
What's the NASA sales team?
Does the company own my side project?
What job openings do we have?
How does compensation work?
Sources
+2
Type your question...

How do I submit timesheets?

To submit a timesheet, log in and navigate to your homepage. Click the timesheets button then select the week to enter. When finished for the week, click the submit button.

Relevant moments

Logging in to your portal 03:50

Homepage tour 07:23

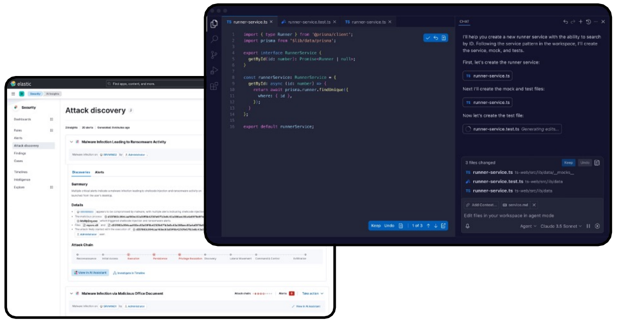
**John**

What features have we released in the last 2 months?

**Aura**

Over the last 2 months, we've released the following features:

1. **Auto-Scaling:** Our Cloud Computing platform now offers auto-scaling capabilities, ensuring optimal resource allocation based on demand.
2. **Enhanced Security:** We've introduced advanced security features, including multi-factor authentication and threat detection, to safeguard your data.



Attack discovery

2 insights | 20 alerts | Generated: 9 minutes ago

Malware Infection Leading to Ransomware Activity

Summary

Details

Attack Chain

Malware Infection via Malicious Office Document

```
import { type Runner } from '@prisma/client';
import prisma from '$lib/data/prisma';

export interface RunnerService {
  getById(id: number): Promise<Runner | null>;
}

const runnerService: RunnerService = {
  getById: async (id: number) => {
    return await prisma.runner.findUnique({
      where: { id },
    });
  }
};

export default runnerService;
```

I'll help you create a new runner service with the ability to search by ID. Following the service pattern in the workspace, I'll create the service, mock, and tests.

First, let's create the runner service:

runner-service.ts

Next I'll create the mock and test files:

runner-service.ts

Now let's create the test file:

runner-service.test.ts Generating edits...

3 files changed

runner-service.ts ts-web/src/lib/data/__mocks__

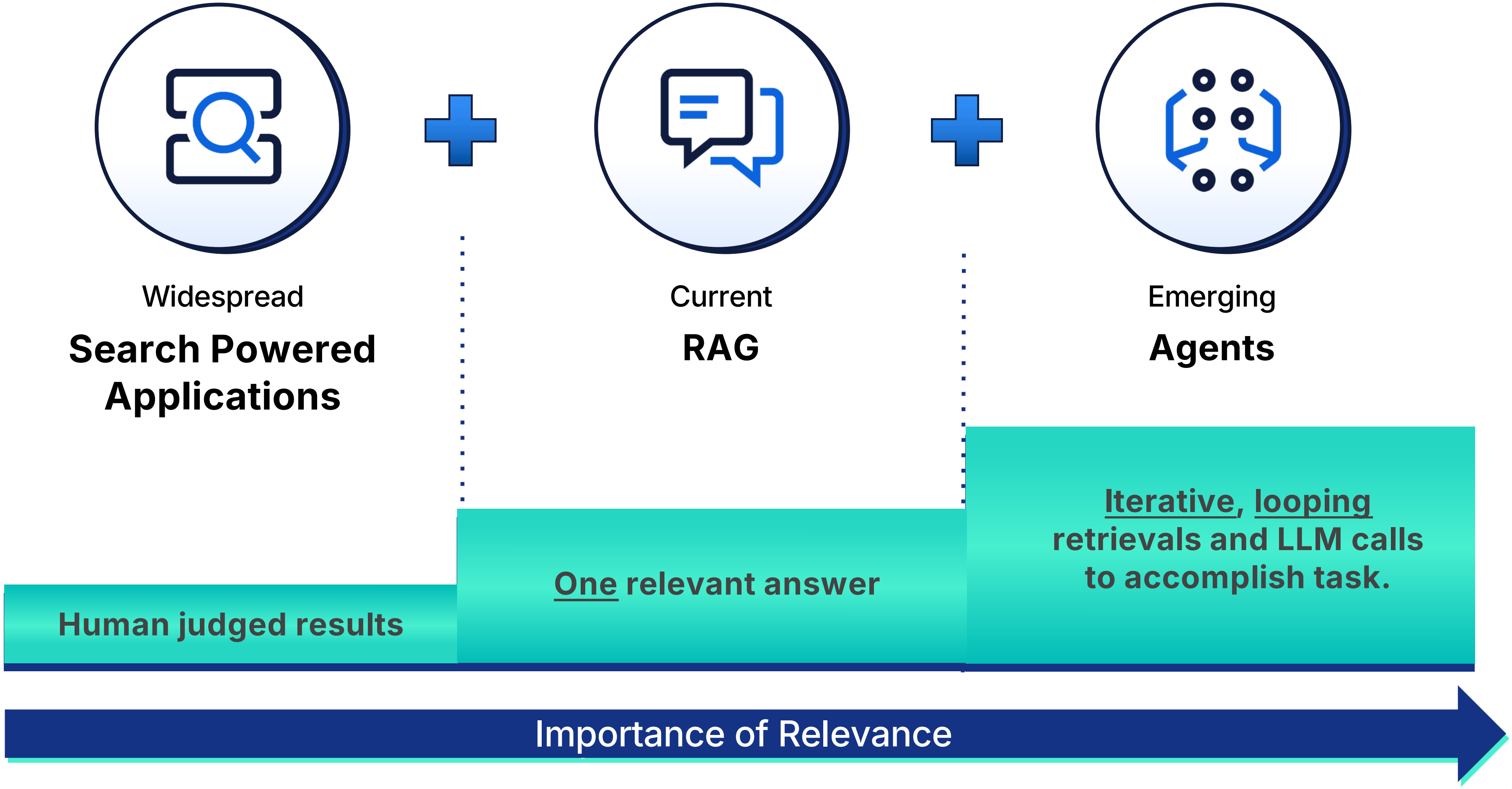
runner-service.test.ts ts-web/src/lib/data

runner-service.ts ts-web/src/lib/data

Edit files in your workspace in agent mode

elastic

# Search relevance and context has never been more important



Widespread

**Search Powered Applications**

Current

**RAG**

Emerging

**Agents**

Human judged results

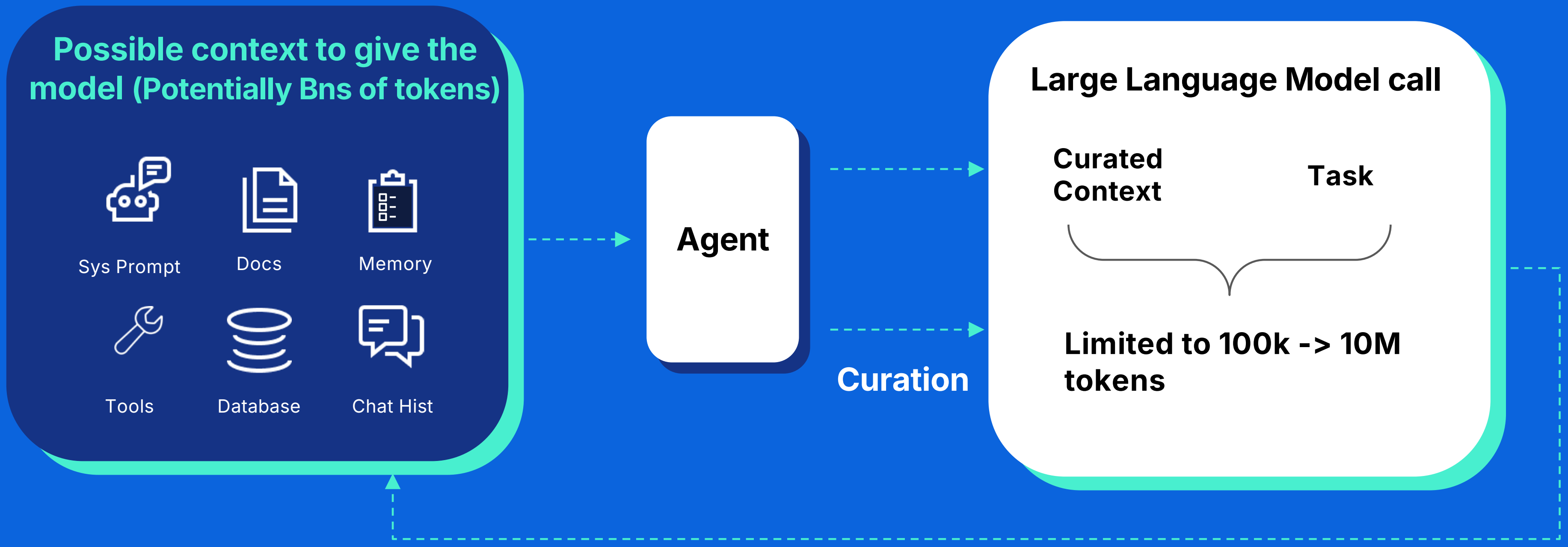
One relevant answer

Iterative, looping retrievals and LLM calls to accomplish task.

Importance of Relevance

elastic

# Improving Agent outcomes through Context Engineering



Possible context to give the model (Potentially Bns of tokens)

Sys Prompt

Docs

Memory

Tools

Database

Chat Hist

Agent

Curation

Large Language Model call

Curated Context

Task

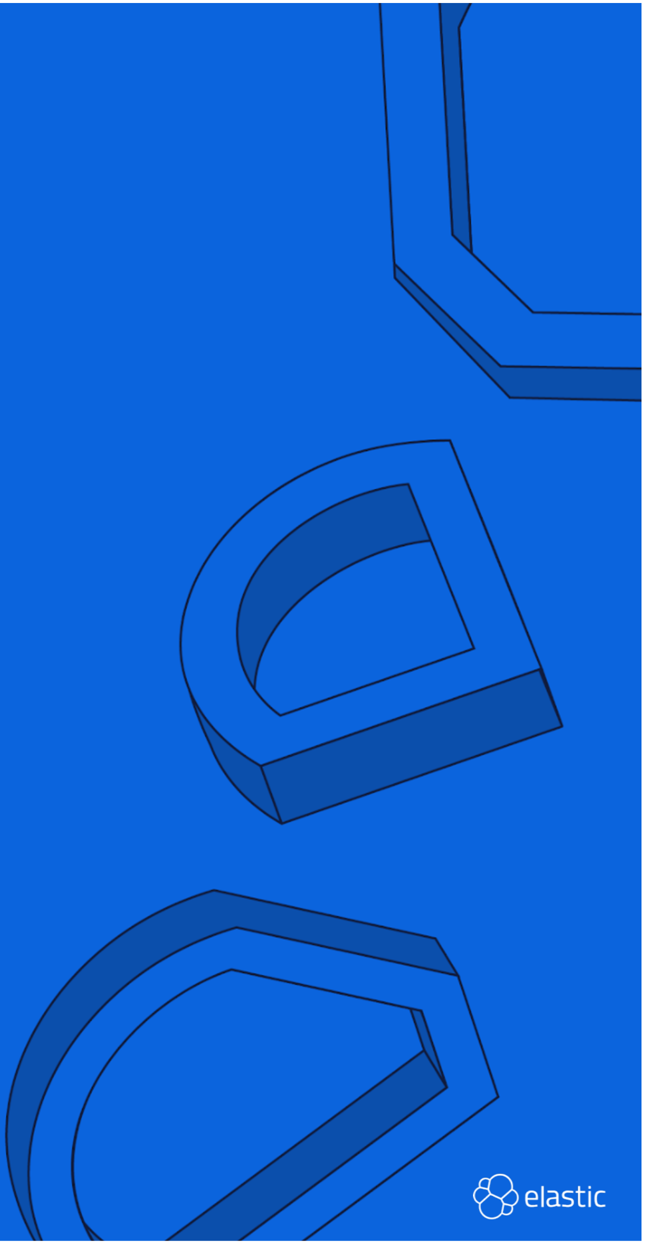
Limited to 100k -> 10M tokens

AI experiences depend on a limited LLM context window.

Dynamically delivering the right data for the LLM’s task is called **Context Engineering**.

elastic

# Agentic AI



# Agentic AI

- **Agentic AI** refers to AI systems that can act autonomously to achieve goals, making decisions and taking actions without direct human intervention
- An **agent** is usually a software that makes decisions and takes actions often by calling tools or APIs in a loop, based on its current understanding of the world and its objective
- Key features of an agent:
  - **Goal-oriented**
  - **Autonomous**
  - **Interactive**
  - **Iterative**: reason → act → observe → repeat
  - **Memory** (optional)

# Agent workflow example

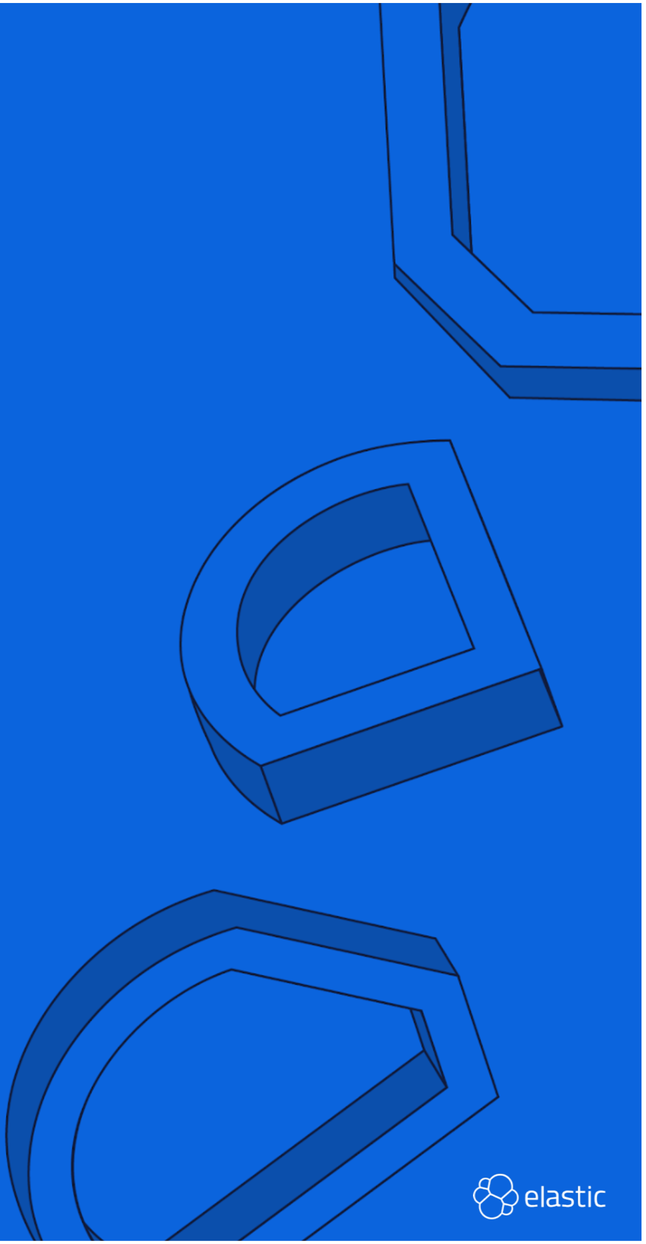
- **Goal:** Book a flight to Paris under $500.
  - **Plan (reason)**: “Check available flights.”
  - **Act**: Calls search_flights(to="Paris", max_price=500)
  - **Observe**: Gets a list of flights.
  - **Reason**: “There’s a flight on Tuesday under budget. Book it!”
  - **Act**: Calls book_flight(flight_id)
  - **Finish**: Reports the result back to the user.
- The agent monitors its own process, deciding what to do next based on each result

# Limitations

- **Execution time:** not really fast since it requires many steps
- **Expensive**: agents typically use about 4x more tokens than chat interactions, multi-agent use about 15x more tokens than chat
- **Complexity**: multi-agent are good with parallel tasks but not so good to manage many dependencies between agents*
- **Unpredictability**: agent interactions can lead to unexpected outcomes. Needs of Guardrail systems.
- **Evaluation**: it's difficult to measure the collective success or alignment of multiple agents

* How we built our multi-agent research system, Anthropic

# Tool calling



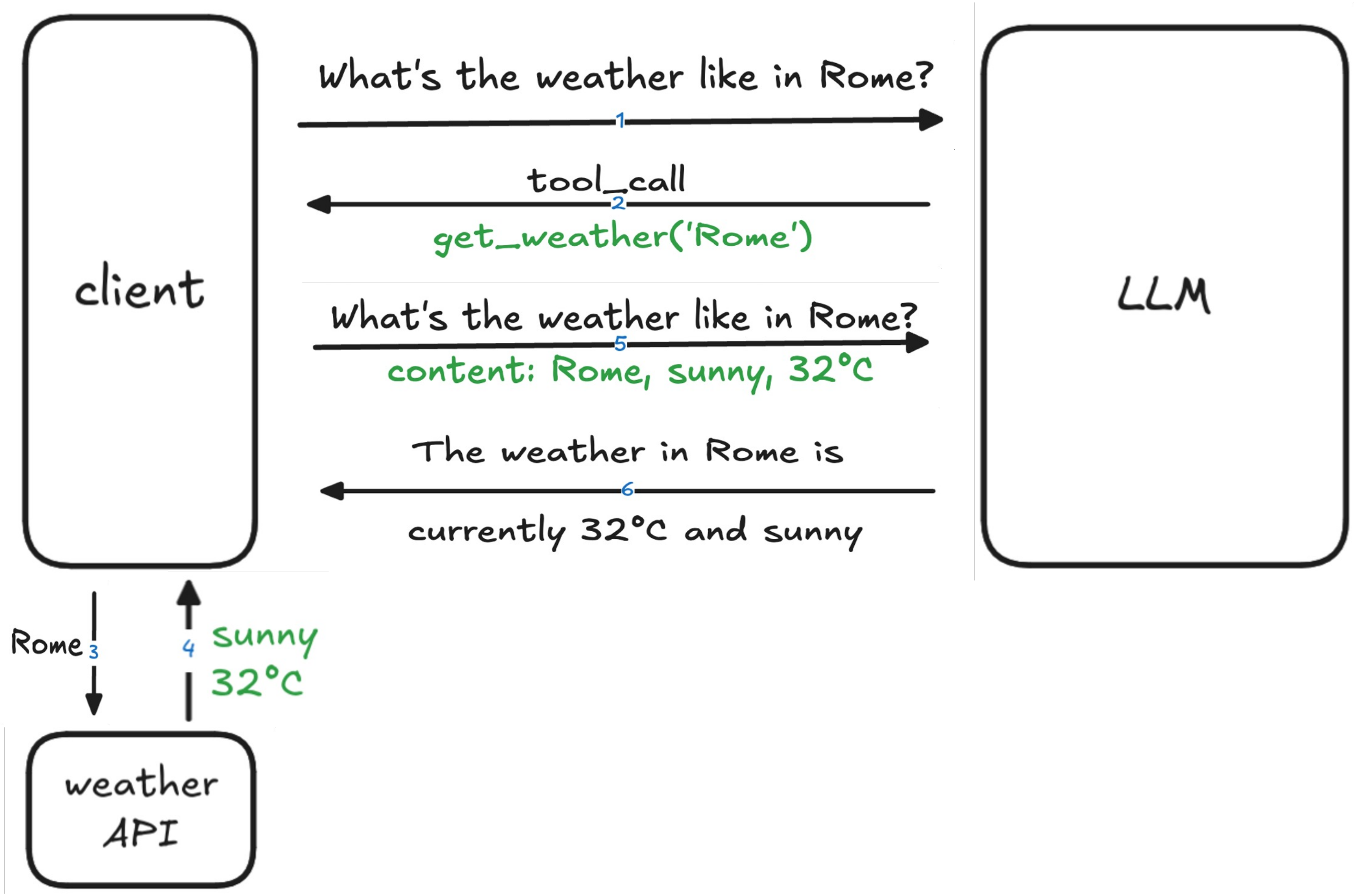
# History of tool calling

- **Tool calling** (or function calling) is an emerging property in LLM
- Relevant papers that investigated the topic:
    - Nakano et al., WebGPT: Browser-assisted question-answering with human feedback, OpenAI, 2022
    - Timo Schick et al., Toolformer: Language Models Can Teach Themselves to Use Tools, Meta AI Research, 2023
    - Function calling and other API updates, OpenAI, June 13, 2023

# Tool calling (or Function calling)

- Tool calling is the ability of LLM to recognize the need to execute external functions (tools) as part of its reasoning process
- The LLM recognizes when it needs of additional information or actions and request the usage of tools (preparing the generation a function call in JSON function)
- The client is responsible for executing the function call (not the LLM) and this step is usually monitored by a human

# A diagram of Tool calling



elastic

# Tool calling in OpenAI

```
POST  https://api.openai.com/v1/chat/completions
{
   "model": "gpt-4.1",
   "messages": [
      {
         "role": "user",
         "content": "What is the weather like in Rome today?"
      }
   ],
   "tools": [
      {
         "type": "function",
         "function": {
            "name": "get_weather",
            "description": "Get current temperature for a given location.",
            "parameters": {
               "type": "object",
               "properties": {
                  "location": {
                     "type": "string",
                     "description": "City and country e.g. Rome, Italy"
                  }
               },
               "required": [
                  "location"
               ],
               "additionalProperties": false
            },
            "strict": true
         }
      }
   ]
}
```

## Response

```
[{
   "id": "call_12345xyz",
   "type": "function",
   "function": {
      "name": "get_weather",
      "arguments": "{\"location\":\”Rome, Italy\"}"
   }
}]
```

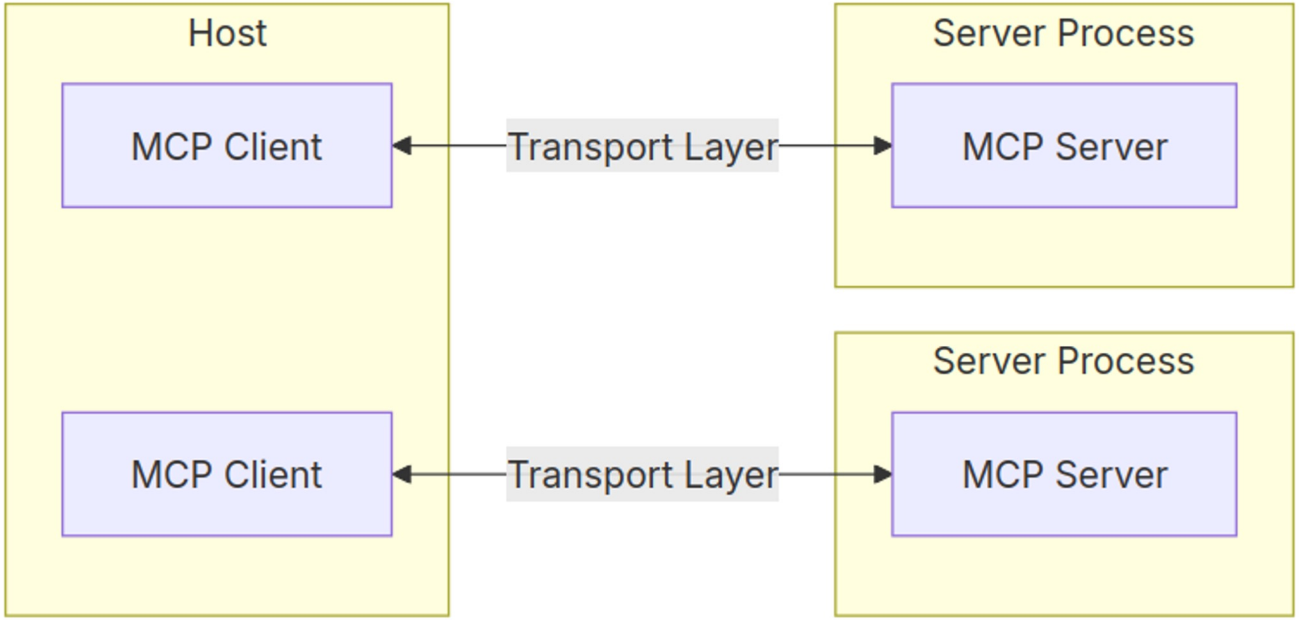
# Model Context Protocol

# Model Context Protocol

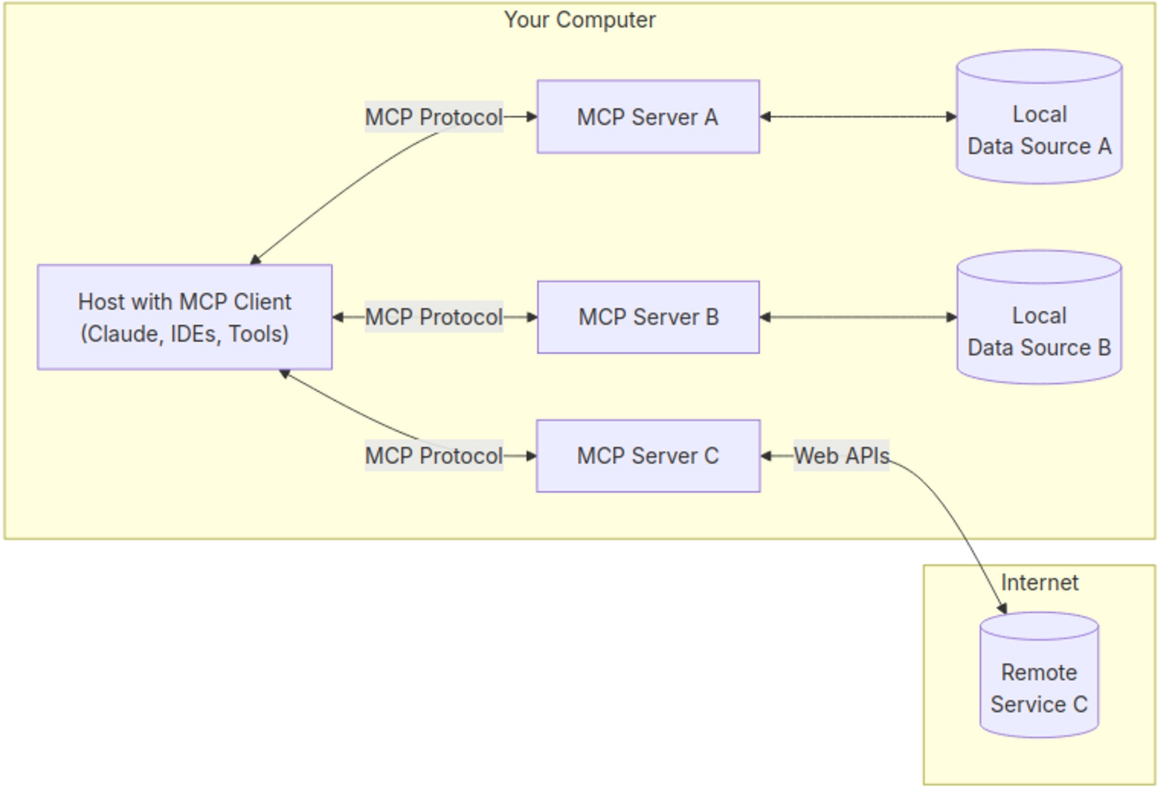
- Model Context Protocol (MCP) is an open protocol that standardizes how applications provide context to LLMs
- MCP provides a standardized way to connect AI models to different data sources and tools
- Client/server architecture

# Core architecture

- MCP follows client-server architecture
- Transport layer: Stdio, Streamable HTTP
- All transport use JSON-RPC 2.0 to exchange messages



# Example: multiple MCP servers



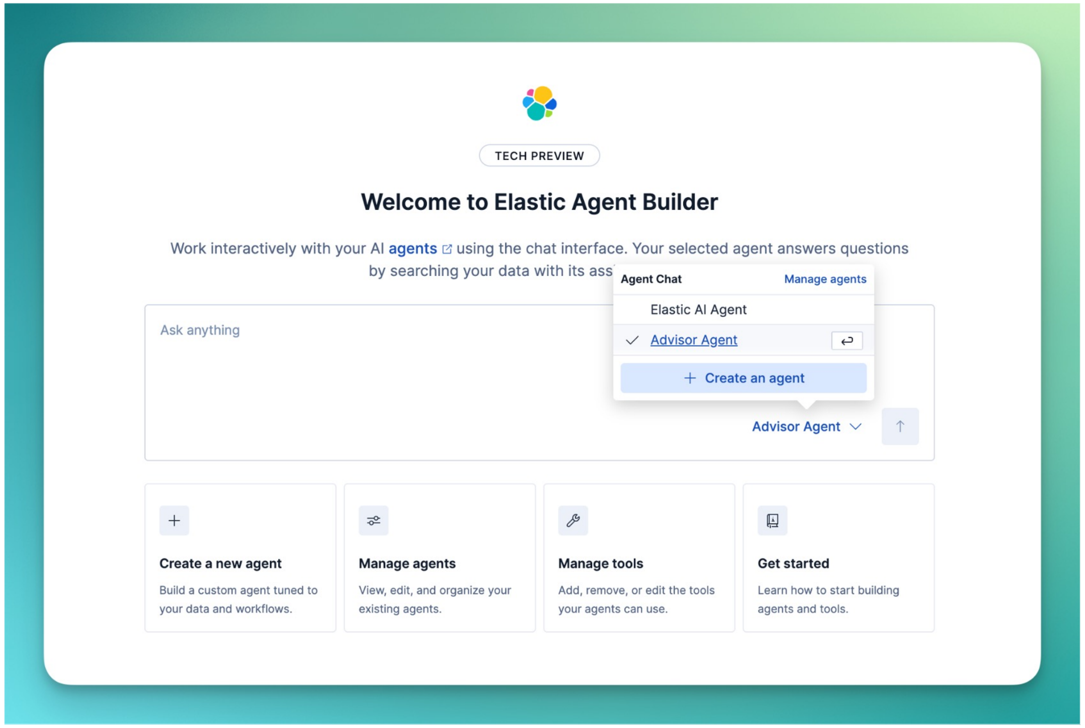
elastic

Elastic 9.2 Tech Preview

# Meet **Elastic Agent Builder**

Develop custom AI agents and chat experiences in minutes with Elasticsearch leading relevance and new agent-building capabilities

- ➔ **Natively chat with any data in Elastic** using a built-in agent
- ➔ **Build custom AI agents** that achieve higher accuracy, relevance, and efficiency based on the power of hybrid search
- ➔ **Create powerful tools:** Give your agent new powers. Build custom tools with the full power of ES|QL for fine-grained control over data, relevance, and security.
- ➔ **Expose** your data by hosting **MCP and A2A** compatible interfaces to your AI ecosystem.



TECH PREVIEW

### Welcome to Elastic Agent Builder

Work interactively with your AI agents using the chat interface. Your selected agent answers questions by searching your data with its ass

**Create a new agent**
Build a custom agent tuned to your data and workflows.

**Manage agents**
View, edit, and organize your existing agents.

**Manage tools**
Add, remove, or edit the tools your agents can use.

**Get started**
Learn how to start building agents and tools.

elastic

# Conversational chat can help you ...

Improve technical support resolution times with Agents to aid investigations

Improve sales reporting and analysis with Agents to summarize and deliver insights

Understand and create policies (e.g. HR, IT, Legal) based on existing and usage (e.g. past searches/conversations)

Investigate purchase and product trends to make better recommendations

Create intelligent product or content recommendations based on catalog and past searches, interactions

Any other internal knowledge worker use case...

Ensure compliance across unstructured/structured data sources (e.g. KYC)

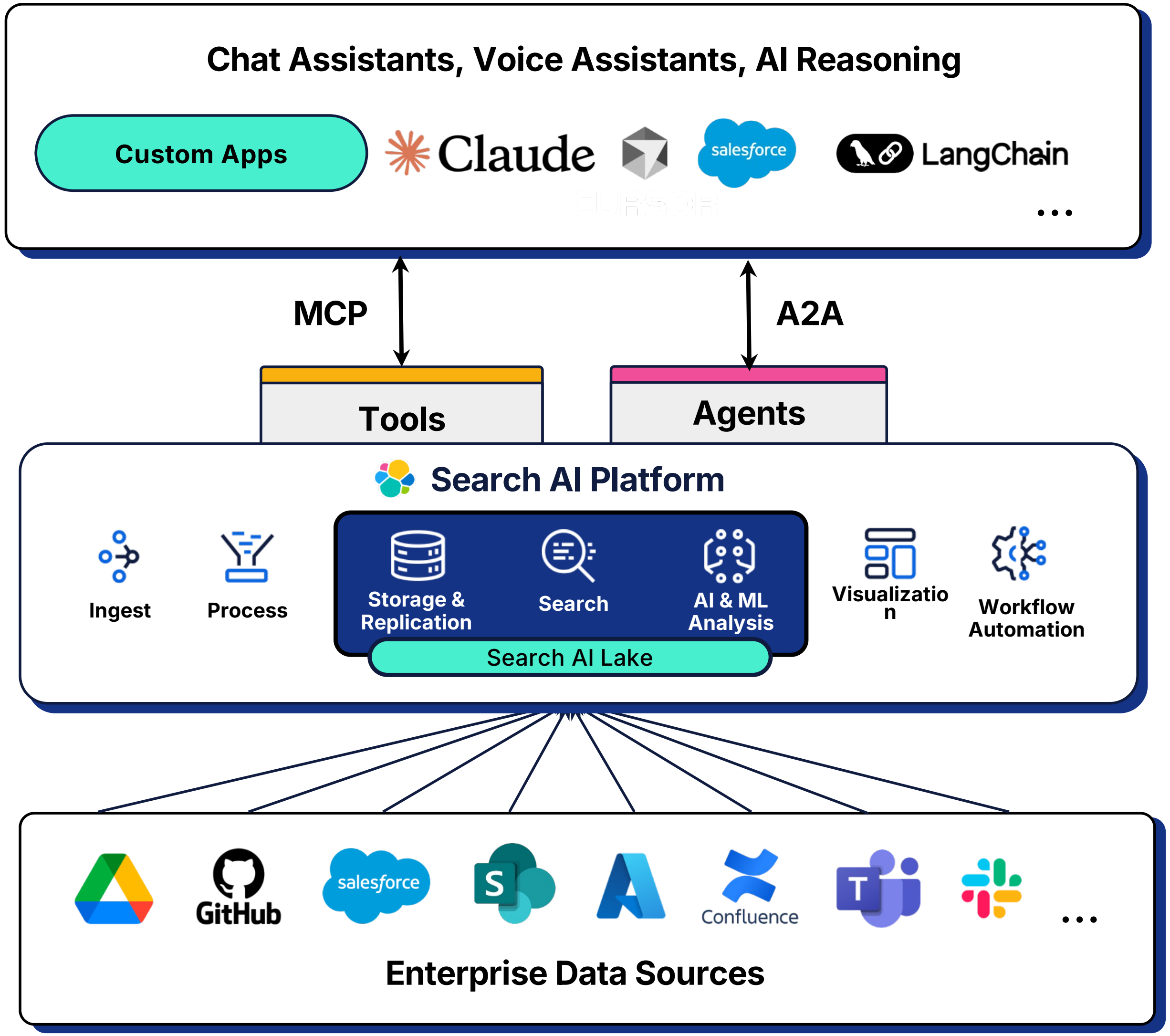
# One unified platform to build custom AI agents fast

**AI Native Experiences**

**Powered by…**
Tools & Agents

**Enabled by…**
The Platform

**Built on…**
Enterprise Data



Chat Assistants, Voice Assistants, AI Reasoning

Custom Apps

Claude

salesforce

LangChain

…

MCP

A2A

Tools

Agents

Search AI Platform

Ingest

Process

Storage & Replication

Search

AI & ML Analysis

Visualizatio n

Workflow Automation

Search AI Lake

GitHub

salesforce

Confluence

…

Enterprise Data Sources

# Quickly build custom agents that utilize all your data powered by Elasticsearch context

Build on the best Vector Database technology

***Speed, Scale, Efficiency, Relevance***

Data Relevance built to match the intent of users

***Hybrid Search, Machine Learning, Small Language Models***

Unify Agent Data layer with Enterprise quality

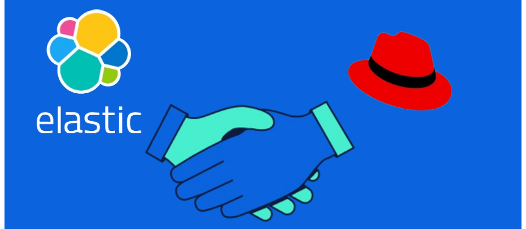
***Data Security, Multi-Cloud***

# Elasticsearch (ECK) Operator

✓ Certified

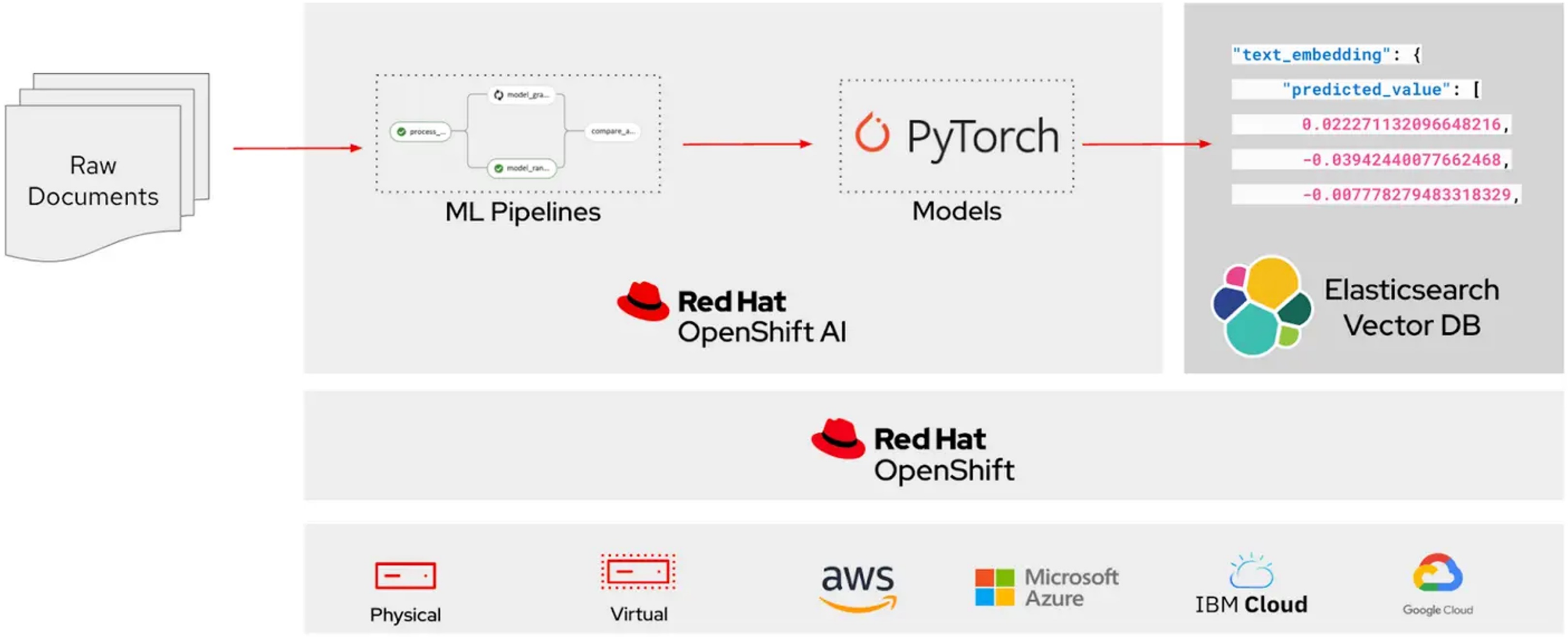
ECK into Red Hat Catalog

## Elastic
## &
## Red Hat



- Easily manage the Elastic Stack on Openshift
- Secure by Default
- Advanced Topology Support
- Snapshot scheduling and keystore support
- Cross-cluster search and cross-cluster replication
- Store local, search Global

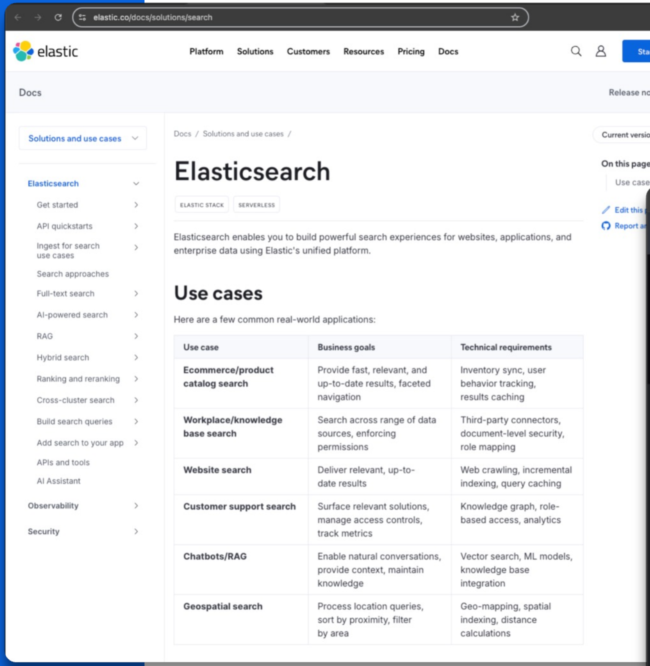
# Red Hat Openshift AI & Elasticsearch Vector DB



```
"text_embedding": {
    "predicted_value": [
        0.022271132096648216,
        -0.03942440077662468,
        -0.007778279483318329,
```

# Want to
# learn more?

Elastic Docs | Quick Start



Elasticsearch

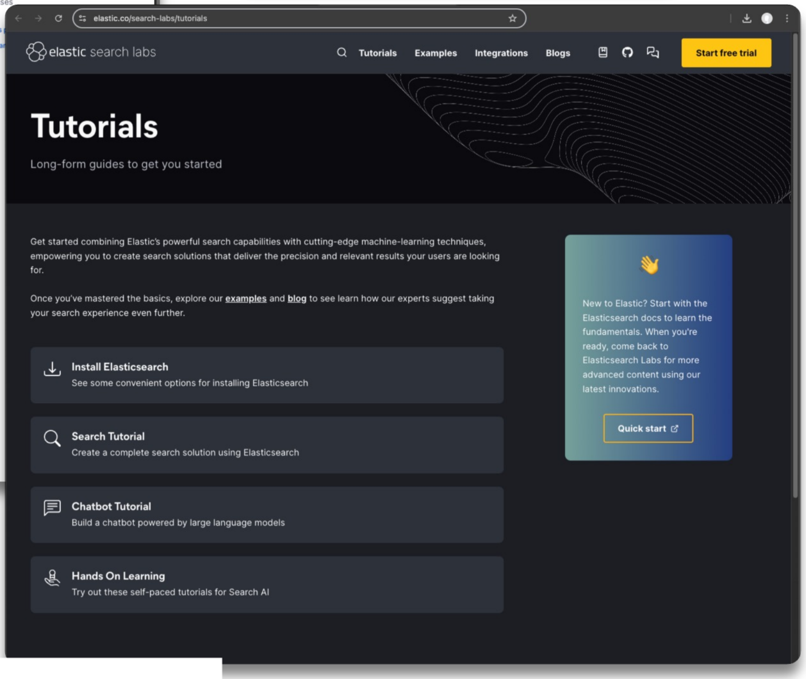
Elasticsearch enables you to build powerful search experiences for websites, applications, and enterprise data using Elastic's unified platform.

Use cases

Here are a few common real-world applications:

| Use case | Business goals | Technical requirements |
|---|---|---|
| **Ecommerce/product catalog search** | Provide fast, relevant, and up-to-date results, faceted navigation | Inventory sync, user behavior tracking, results caching |
| **Workplace/knowledge base search** | Search across range of data sources, enforcing permissions | Third-party connectors, document-level security, role mapping |
| **Website search** | Deliver relevant, up-to-date results | Web crawling, incremental indexing, query caching |
| **Customer support search** | Surface relevant solutions, manage access controls, track metrics | Knowledge graph, role-based access, analytics |
| **Chatbots/RAG** | Enable natural conversations, provide context, maintain knowledge | Vector search, ML models, knowledge base integration |
| **Geospatial search** | Process location queries, sort by proximity, filter by area | Geo-mapping, spatial indexing, distance calculations |

Elasticsearch Lab Tutorials



Tutorials

Long-form guides to get you started

Get started combining Elastic's powerful search capabilities with cutting-edge machine-learning techniques, empowering you to create search solutions that deliver the precision and relevant results your users are looking for.

Once you've mastered the basics, explore our examples and blog to see learn how our experts suggest taking your search experience even further.

**Install Elasticsearch**
See some convenient options for installing Elasticsearch

**Search Tutorial**
Create a complete search solution using Elasticsearch

**Chatbot Tutorial**
Build a chatbot powered by large language models

**Hands On Learning**
Try out these self-paced tutorials for Search AI

New to Elastic? Start with the Elasticsearch docs to learn the fundamentals. When you're ready, come back to Elasticsearch Labs for more advanced content using our latest innovations.

Elasticsearch Hands-On Course

## Elastic Agent Builder - Tools, Agents, and MCP

In this hands-on course, learn how to create custom AI tools from your own business logic, build specialized agents to accurately chat with your data, and integrate them into external applications via the built-in MCP server.

# Thank You

elastic | The Search AI Company

Red Hat Summit

# Connect

in linkedin.com/company/red-hat

f facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat