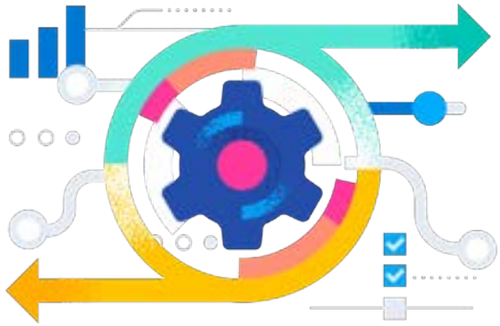
elastic

Generative AI and Elastic

Unleash the power of your data and safeguard privacy

aws PARTNER

Today, artificial intelligence (AI) is at the center of nearly every conversation about the future of technology in business. As models improve and the infrastructure to support machine learning (ML) at massive scale grows more robust, it's only going to become more necessary to utilize AI to accelerate innovation, product development, cybersecurity, and a host of other operations.



While AI provides answers to critical business challenges, it also raises questions. Which provider do you partner with? What data do you unlock? How do you manage privacy and other security concerns? How do you keep up with an accelerated rate of change while protecting your business from the inevitable pitfalls of any new technology?

At Elastic, we've built AI into each of our solution areas, making it easy to integrate powerful large language models (LLMs) into your operations. We're building generative AI solutions with a privacy-first approach to make data both available and secure. With this approach, you can help your teams fully leverage generative AI's potential—and do it faster—while safeguarding private data and avoiding AI hallucinations created by false or inaccurate data.

With Elastic Cloud on Amazon Web Services (AWS), you combine Elastic's superior search capabilities and the power of AWS infrastructure and AI/ML-managed services, enhancing the speed, scale, and relevance at which you can innovate with generative AI.

The building blocks of generative AI

Across every organization, there are thousands of potential use cases for generative AI, from HR and finance to product development and customer service. To develop a successful generative AI solution for any use case, you need to deliver accurate, meaningful, and relevant results.

This takes four basic building blocks.

Knowledge base

To make results specific to your organization, your solution needs access to the entire knowledge base in your private store, which can exist in structured datasets or unstructured forms, such as images, videos, and documents.

Natural language processing

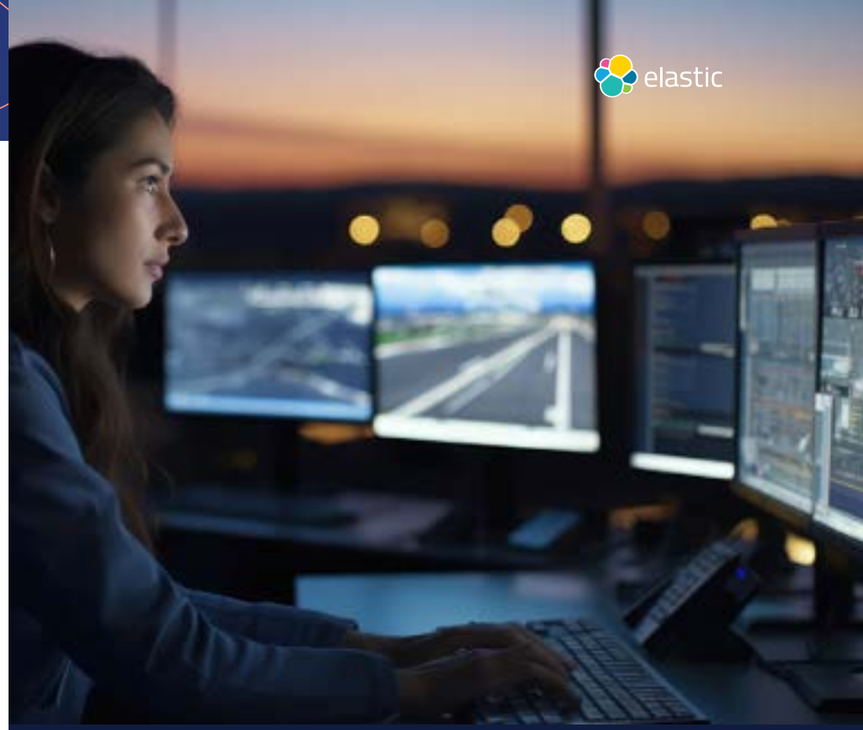
Natural language processing (NLP) capabilities allow your app to understand language, including the underlying meaning and sentiment of text. NLP also enables the solution to generate coherent and contextually relevant content.

Search

The relevance and quality of the content your generative AI solution creates depends on the power of the search tool it uses to gather data. The stronger your search tool, the better the output of your solution.

Large language models

LLMs digest and contextualize data, then render it in natural language to produce meaningful, conversational content based on either public or proprietary ML models.



Imagine…

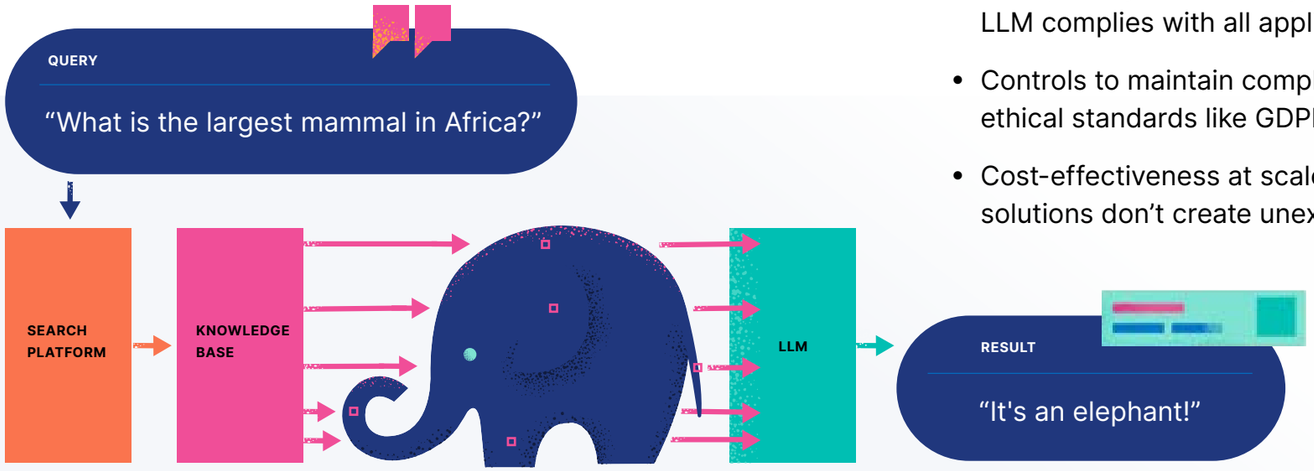
In Elastic's World of Possibility, a transportation authority can use generative AI to turn data related to everything from people to packages into easily accessible insights, transforming the future of movement in a city of six million people.

more →

Search | The heart and soul of generative AI

A famous parable from India tells of a group of blind men who come across an elephant and try to discern its nature by touching it. What is this massive thing before them? Each touches a different part, leading them to perceive and describe the elephant in vastly different ways. While the story is often used to illustrate how people can become fixated on their own versions of the truth, it's also a powerful illustration of the risk of drawing conclusions from too little data.

The search platform delivering data to your LLM defines the accuracy and quality of your solution's results. It's what allows you to see "the whole elephant" in your data.



What should you look for in a search platform?

- A comprehensive, diverse, updated index to enrich the model with highly relevant data.
- Advanced search features, including NLP, semantic search, and contextual understanding to enhance the LLM's ability to process complex prompts.
- Fast, efficient search capabilities to deliver an exceptional user experience.
- Scalability to handle increasing data volumes and user demand.
- Data privacy and security controls to ensure your LLM complies with all applicable standards.
- Controls to maintain compliance with legal and ethical standards like GDPR and CCPA.
- Cost-effectiveness at scale so generative AI solutions don't create unexpected expense.

The generative AI capabilities in Elastic Cloud are built on Elasticsearch, the gold standard in enterprise search.

6 billion requests every day

3,500+ unique contributors

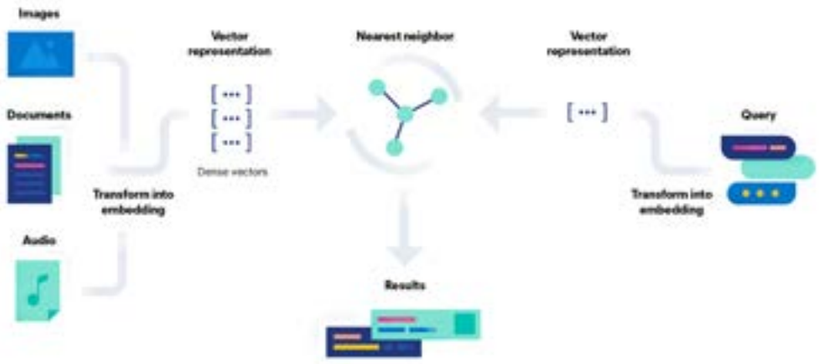
183,000 GitHub stars

4.2 billion downloads

Vector search | Gain insights from all your data

Not every customer will know the right words to use in a prompt to generate the specific results they're looking for. And not all data will be clean and structured. Up to 90 percent of the new data businesses produce is unstructured—messy, complex datasets that hold immense potential.

Vector search leverages ML to transform the meaning and context of unstructured data into a numeric representation, then finds similar data using approximate nearest neighbor (ANN) algorithms. This allows your users to search for what they mean—and get answers quickly based on similarity search.



VECTOR SEARCH ACCELERATES RESULTS

Vector embedding

Numeric expressions of data and its context are stored in high-dimensional vectors, generated by models trained on millions of examples to deliver more relevant and accurate results.

Similarity score

When data and documents are similar, so are their vectors. Indexing queries and documents with vector embeddings reveals similar documents as the nearest neighbors of the query.

ANN algorithm

Using ANN instead of traditional nearest neighbor algorithms saves time and computational resources by executing efficiently in high-dimensional embedding spaces at scale.

Imagine…

In Elastic's World of Possibility, a communications company can break down silos and fully empower its teams to design a global communications infrastructure for a dynamic future building generative AI solutions with Elasticsearch.

more →

What makes Elasticsearch different?

Vector search works fast, but ANN algorithms aren't as accurate as traditional nearest neighbor (kNN) algorithms. In the same way that a fast answer might not be the best answer in a conversation, using ANN algorithms alone can generate results with less robust context. Elasticsearch solves this problem by combining ANN and kNN algorithms to achieve both speed and accuracy. And incorporating Best Match 25 (BM25) similarity ranking enhances the relevance of the unstructured data returned by a query.

ANN algorithm

Conserve resources and generate results fast by scoring similarity exactly.

RICH CONTEXT FAST

kNN algorithm

Improve accuracy by also scoring similarity with more robust contextual relevance.

In addition, Elasticsearch allows you to do more with a search, with faceting and role-based access control (RBAC) integrated into queries.

Faceting

Generate more accurate results by applying filters based on a faceted classification of data.

RBAC

Protect sensitive data by allowing queries access only to data appropriate to a user's role.

elastic

RBAC is fundamental to Elastic's privacy-first approach to generative AI.

By making access control an inherent and automated function of the search platform, Elasticsearch throws the door to innovation wide open—while simultaneously ensuring data privacy, security, and compliance with legal and ethical standards.

Streamline training with Elasticsearch Relevance Engine

For many organizations, the primary obstacle to innovating with generative AI is the expertise needed to create your own LLMs. Elastic removes the friction from this process with the Elasticsearch Relevance Engine™ (ESRE), which allows you to integrate with external LLMs, implement hybrid search, and train models to deliver the results you need without deep AI/ML expertise.

ESRE accelerates developing generative AI solutions by putting the tools you need to customize a model at your fingertips.

Vector database

Capture the meaning and context of your unstructured data by creating embeddings for a full vector search experience at scale.

Elastic Learned Sparse EncodeR (ELSER)

Gain highly relevant semantic search without domain adaptation to expand queries with related keywords and relevance scores.

Retrieval Augmented Generation (RAG)

Give LLMs business-specific data using Elasticsearch for high-relevance context windows to train the model for your specific needs.

RRF hybrid ranking

Combine document rankings from multiple retrieval systems to tune search results from multiple retrievers with less effort.

Bring your own transformer models

Use your own models or upload pretrained third-party models, with support for a variety of architectures.

Data integrations and ingestion libraries

Use familiar ingestion tools and a vast array of data integrations to maximize the data available to your solution.

Search + generative AI | Real-time search analytics

Building on the speed and quality of Elasticsearch, generative AI solutions developed in Elastic Cloud allow you to generate real-time search analytics for discovery, prediction, and prescriptive guidance across the entire data estate.

Security

Operationalize your security strategy using predictive analytics, cloud monitoring, and more.

Observability

Resolve problems faster and discover new opportunities with observability that enables automation and anomaly detection.

Search analytics

Create powerful real-time search experiences for employees and customers.

For our customers, that means deeper insights that paint a brighter, bolder, more actionable picture of where the organization is and where it's going, all easily accessible through a sophisticated, conversational, and well-governed generative AI framework.

Walmart

Walmart protects customers and combats fraud by using real-time search and Elastic Security to identify instances of fraud in real time, based on billions of metadata records ingested into Elastic Cloud.

COMCAST

Comcast gains a single, real-time view across 50,000 software builds and 400 terabytes of telemetry data ingested every day. Elastic Observability aggregates, correlates, and inspects the data from about 70 tenants across its infrastructure.

Imagine…

In Elastic's World of Possibility, a medical network, with a comprehensive view of all its data, can connect every aspect of a patient's health and experience to improve preventative care and increase the patient's long-term well-being.

more →

Elastic on AWS | Elasticsearch meets Amazon Bedrock

Amazon has been innovating with AI/ML for more than 25 years, from the models behind Amazon's recommendation engine and vast logistics network to the services supporting Amazon and AWS customers around the world.

At its core, generative AI is powered by foundation models (FMs), pretrained on massive data sets, that can be customized using your data for domain-specific tasks.

Amazon Bedrock, a fully managed service for building generative AI solutions, provides a wide range of high-performing FMs, including leading open LLMs, such as Mistral, and closed LLMs, such as Anthropic, Claude2, and Cohere. Using these models, you can experiment with and customize your data using techniques like fine-tuning and Retrieval Augmented Generation (RAG).

Amazon Bedrock, Amazon Sagemaker, and other AWS AI solutions integrate with Elastic Cloud on AWS, allowing you to combine the power of Elasticsearch with managed services that deliver:

- Flexibility and choice of FMs
- Secure customization
- Cost-effective infrastructure
- High-performance, low-cost ML
- The easiest way to build custom ML solutions and apply FMs

AWS has purpose-built processors to deliver the performance needed for a quality generative AI experience while controlling costs.

Compared to comparable Amazon EC2 instances:

- **AWS Trainium** delivers up to **50 percent savings** on training costs.
- **AWS Inferentia2** delivers up to **40 percent better price performance**.

Elastic

Accelerates generative AI solution development with search, seamlessly integrates security with the Elastic AI assistant, and boosts productivity with built-in generative AI capabilities across Elastic Cloud.

Amazon Bedrock

Streamlines generative AI solution development with powerful FMs, training solutions, and infrastructure to support generative AI at scale.

Amazon Sagemaker

Allows you to build, train, and apply ML models at scale, while supporting data governance with simplified access control and transparency over your ML projects.

The results

Your generative AI solutions perform better, with relevant results.

Your users get what they need, when they need it, providing an exceptional user experience.

How will generative AI transform your business?

The time for generative AI has arrived, transforming the way businesses around the world interact with their data.

With Elastic on AWS, it's now easier than ever to embrace the full potential of generative AI in your business, with tools, infrastructure, and managed services readily available to support your journey.

Are you ready to imagine?

Investigate what's possible in Elastic's World of Possibility.

→

Dive deeper into what's possible with our webinar, Generative AI and Elastic Maximize Data Value.

→